cemmap

centre for microdata methods and practice

Consistent noisy independent component analysis

Stéphane Bonhomme
Jean-Marc Robin

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP4/08

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

An ESRC Research Centre

Consistent Noisy Independent Component Analysis

Stéphane Bonhomme[1]
CEMFI, Madrid

Jean-Marc Robin
Paris School of Economics,
Université Paris 1,
and University College London

October 2007

[1]**Corresponding author:** CEMFI, Casado del Alisal, 5, 28014 Madrid, Spain. E-mail: bonhomme@cemfi.es

Abstract

We study linear factor models under the assumptions that factors are mutually independent and independent of errors, and errors can be correlated to some extent. Under factor non-Gaussianity, second to fourth-order moments are shown to yield full identification of the matrix of factor loadings. We develop a simple algorithm to estimate the matrix of factor loadings from these moments. We run Monte Carlo simulations and apply our methodology to British data on cognitive test scores.

JEL codes: C14.

Keywords: Independent Component Analysis, Factor Analysis, high-order moments, noisy ICA.

1 Introduction

A linear factor model relates a vector of L measurements to a vector of K unobserved sources, or factors, *via* a linear relationship:

$$Y = \Lambda X + U, \tag{1}$$

where Λ is L-by-K matrix of parameters (factor loadings) and U is a vector of L errors. In Factor Analysis (FA) sources are assumed orthogonal, and Λ is identified up to a rotation (Anderson and Rubin, 1956). Independent Component Analysis (ICA) strengthens the orthogonality assumption, and assume that all components of X and U are independent. Then, if $K \leq L(L-1)/2$ and factors are *not* normally distributed, Λ is generically identified up to sign and permutation normalizations (Comon, 1994, Eriksson and Koivunen, 2003). In the past ten years, ICA has become the standard approach to source separation, with numerous applications to signal processing, telecommunications, and medical imaging (Hyvärinen, Karunen and Oja, 2001).

In those fields where factor analysis is widely used, such as finance, macroeconomics and psychometrics, factor independence may be a natural assumption. For example, in Ross's (1976) Arbitrage Pricing Theory expected returns on assets are modelled as linear combinations of independent factors. Other examples are structural VAR models (e.g., Blanchard and Quah, 1989) and dynamic factor models (e.g., Forni and Reichlin, 1998), if one assumes that the underlying structural innovations are independent instead of simply uncorrelated. If the data are sufficiently non-normal, the higher-order orthogonality conditions implied by independence can then be used to allow for more factors and, if needed, to identify the rotation.

The particular application that we consider in this paper refers to psychometrics. We use an independent factor model to analyze cognitive test scores. Factor analysis has been widely used for this purpose, since the pioneering work of Spearman (1904) and

Thurstone (1947). Recent microeconometric studies also use independent multi-factor models to reveal various dimensions of ability; see for instance Carneiro, Hansen and Heckman (2003) and Heckman, Stixrud and Urzua (2006). In these papers second-order information is sufficient to identify factor loadings because *ex-ante* restrictions are made on factor loadings, each unobserved factor having at least two specific measurements.

In comparison, we exploit the identifying content of the independence assumption. Studying the so-called Thurstone's box problem, Jennrich and Trendafilov (2005) have recently shown that rotating factor loadings towards independence rather than using *a priori* restrictions yields interpretable loadings (see also Mooijaart, 1985). Indeed, the most independent rotations are also the ones that maximize factor non-Gaussianity, an idea that closely matches Kaiser's (1958) *Varimax* rotation criterion, widely used in psychometrics (Kano *et al.*, 2003).

As cognitive tests do not perfectly measure ability, the additive noise cannot be neglected in our application. This is generally the case in applications to econometrics, psychometrics and finance, because of the presence of measurement error or specific factors. However, most ICA algorithms do not explicitly allow for noise. Indeed, in ICA applications errors are usually assumed negligible ($U \approx 0$), and noise-free methods are applied hoping that the signal-to-noise ratio will be high enough for the bias due to neglecting errors to be small (Cardoso and Pham, 2004). The methodological contribution of this paper is to fill this gap in the literature, and to provide a close substitute to noise-free ICA algorithms that remains consistent in the presence of noise.

Our approach builds on the ICA literature. In the noise-free case, several efficient ICA algorithms are currently available to separate up to $K = L$ unobserved factors, FastICA (Hyvärinen and Oja, 1997) and JADE (Cardoso and Souloumiac, 1993) being especially popular. Most of these methods use a two-step approach to estimation (Chen and Bickel, 2005). In the first step (*prewhitening*), the data are transformed so that the covariance

matrix is the identity, e.g. using Principal Component Analysis (PCA). In the second step (*source separation*) the rotation matrix is derived from higher-order information.

Two approaches have been proposed to deal with noisy ICA models. However, these two approaches are not without drawbacks for our purpose. In the first approach (Moulines *et al.*, 1997, Attias, 1999) a flexible parametric model is postulated for factor and error distributions. Maximum Likelihood is often used for estimation, together with the EM algorithm. This requires an appropriate parametric specification, raising e.g. the issue of the number of components in mixture models, as well as computational difficulties.

The second approach relies on a prewhitening step as in noise-free ICA methods, replacing PCA by Probabilistic PCA (Beckmann and Smith, 2004) or FA (Ikeda and Toyama, 2000, Stegeman and Mooijaart, 2007). This approach yields a fast semi-parametric estimation of Λ. Yet, as only second-order moments of the data are used in the prewhitening step, the number K of common factors must be less than the Ledermann bound ($K = (2L + 1 - \sqrt{8L + 1})/2$ if errors are mutually independent) for the procedure to be consistent. Moreover, it only deals with Gaussian errors. If errors are sizeable and the data are highly non-normal, this assumption can be problematic.

We also adopt a semi-parametric, two-step approach. In the first step, second to fourth-order moments of error variables are inferred from a set of linear restrictions, and filtered out from the corresponding data moments. Importantly, unlike the previous literature we use all second to fourth-order data moments in the first step. Then, the second step uses Cardoso and Souloumiac's (1993) JADE algorithm to estimate factor loadings. We call quasi-JADE this two-stage estimation procedure.

Quasi-JADE is consistent whether errors are Gaussian or not, and is almost as fast to run as JADE. An important property of the algorithm is that errors can be correlated to some extent. We show that, if J is the number of mutually independent error pairs, up to

$K = \min\{J, L\}$ factors are generically identified. In the particular case of independent errors, we can thus relax the Ledermann bound and estimate up to L factors. This is because we use higher-order data moments in the prewhitening step of the algorithm.

The algorithm can be applied to cases where factor loadings are restricted *ex-ante*, as in structural VARs. If there are sufficiently many restrictions for the rotation indeterminacy to disappear, factor loadings and error covariances can be jointly estimated from the first estimation step. The benefits of using higher-order information then translate into the possibility of allowing for a richer error structure.

The estimation procedure uses information from second, third and fourth-order moments of the data, while most ICA algorithms assume symmetric factors and discard third-order information. In econometric applications, though, third-order moments can be informative. Following Geary (1942) and Reiersol (1950), a long series of econometric contributions have proposed different ways to combine second and third-order moments to identify factor loadings in the linear *measurement error* (one-factor) model. See e.g. Pal (1980), Dagenais and Dagenais (1997), Lewbel (1997), and Erickson and Whited (2002). The estimator introduced in this paper can be seen as a generalization of this approach to multi-factor structures.

Finally, we also consider the case of *overcomplete* ($K > L$) ICA models with restrictions on factor loadings. Our estimation procedure can be applied iteratively to estimate a model with L unrestricted factors, $L-1$ factors specific to measurements $\{2, ..., L\}$, $L-2$ factors specific to measurements $\{3, ..., L\}$, etc., and one last factor specific to the last two measurements, for a total of $L(L-1)/2$ factors. Estimating overcomplete models is a notoriously difficult problem (e.g., Comon, 2004, for the case $L = 2$, $K = 3$). To our knowledge we are the first to provide a simple, consistent estimation procedure for a large class of overcomplete models.

Sections 2 and 3 present the model, derive the moment restrictions on which identi-

fication and estimation are based and show the identification of the number of factors, error cumulants and factor loadings. In Section 4 we discuss the estimation of the number of factors and factor loadings, and develop the asymptotic distribution theory for JADE, surprisingly missing in the literature. In Section 5, we illustrate the finite-sample properties of our procedure by means of Monte-Carlo simulations, and in Section 6 we apply the method to British data on cognitive test scores. Lastly, Section 7 concludes.

2 Model and moment restrictions

2.1 The model

Let $Y = (Y_1, ..., Y_L)^{\mathrm{T}}$ be a vector of $L \geq 2$ zero-mean, real-valued random variables (measurements), where $^{\mathrm{T}}$ denotes the transpose operator. Let $X = (X_1, ..., X_K)^{\mathrm{T}}$ be a random vector of $K \geq 1$ real valued, non degenerate random variables (factors). Let also $U = (U_1, ..., U_L)^{\mathrm{T}}$ be a vector of L real-valued random variables (errors). An observation sample is a collection of N independent draws of vector Y.

Assumption A1 *There exists a L-by-K matrix of scalar parameters (factor loadings), $\Lambda = [\lambda_{\ell k}]$, such that $Y = \Lambda X + U$, and Λ, X and U satisfy the following conditions:*

1. *$(X^{\mathrm{T}}, U^{\mathrm{T}})^{\mathrm{T}}$ has zero mean and finite moments up to the fourth order.*

2. *The components of X are mutually independent, and independent of those of U.*

3. *The components of X have unitary variance.*

A triple (Λ, X, U), satisfying these assumptions is called a representation.

In the second statement, independence can be replaced by the weaker assumption of zero multivariate cumulants up to the fourth order, as we shall only consider moments up to the fourth order for identification and estimation. The third statement is a normalization condition. If (Λ, X, U) is a representation, then $(\Lambda D^{-1}, DX, U)$ is another

representation for any diagonal matrix D with positive entries on the diagonal. Hence, one may as well normalize the variance of each component of X to unity.

The normalization of the variance of X is not sufficient to grant identification. For any value of K, the number of factors, let us define the set of sign-permutation matrices as the set $\mathcal{S}_K$ of all products DP, where D is a diagonal matrix with diagonal components equal to 1 or -1 and P is a permutation matrix. For given values of L and K, let (Λ, X, U) be a representation. Clearly, for all $S \in \mathcal{S}_K$, $\left(\Lambda S, S^{\mathrm{T}} X, U\right)$ is another representation. We say that the matrix of factor loadings Λ is identified if any equivalent representation $(\widetilde{\Lambda}, \widetilde{X}, \widetilde{U})$ is such that Λ and $\widetilde{\Lambda}$ are equal modulo $\mathcal{S}_K$ (i.e. $\widetilde{\Lambda} = \Lambda S$ for some $S \in \mathcal{S}_K$).

Given the linearity and independence assumptions, working with cumulants is especially convenient. Multivariate cumulants of centered random variables of order 2, 3 and 4 are defined as follows:

$$\begin{aligned}
\mathrm{Cum}\,(Z_1, Z_2) &= \mathbb{E}(Z_1 Z_2), \\
\mathrm{Cum}\,(Z_1, Z_2, Z_3) &= \mathbb{E}(Z_1 Z_2 Z_3), \\
\mathrm{Cum}\,(Z_1, Z_2, Z_3, Z_4) &= \mathbb{E}\left(Z_1 Z_2 Z_3 Z_4\right) - \mathbb{E}(Z_1 Z_2)\mathbb{E}(Z_3 Z_4) - \mathbb{E}(Z_1 Z_3)\mathbb{E}(Z_2 Z_4) \\
&\quad - \mathbb{E}(Z_1 Z_4)\mathbb{E}(Z_2 Z_3).
\end{aligned}$$

To ensure identification we impose the following restrictions on the first cumulants of error variables.

Assumption A2 *There exists a non empty set of indices $\mathcal{J} \subset \{(\ell, m) \in \{1, ..., L\}^2,\ \ell < m\}$ such that, for all $(\ell, m) \in \mathcal{J}$ and all measurement indices i and j, we have:*

$$\mathrm{Cum}\,(U_\ell, U_m) = \mathrm{Cum}\,(U_i, U_\ell, U_m) = \mathrm{Cum}\,(U_i, U_j, U_\ell, U_m) = 0.$$

Most of the ICA literature makes parametric assumptions on errors, usually assuming Gaussianity. However, Davies (2004) points out that error Gaussianity alone is not sufficient to provide identification of factor loadings in a noisy ICA model. For identification, one needs to restrict the dependence between errors, which is what Assumption A2 does.

The following lemma shows that Assumption A2 is satisfied by a broad class of error structures.

Lemma 1 *Let $U = \Pi\varepsilon$, where Π is a L-by-H matrix of scalar parameters, and the components of ε are mutually independent and independent of those of X with finite moments up to the fourth order. Then U satisfies assumption A2, with*

$$\mathcal{J} = \{(\ell, m) \in \{1, ..., L\}^2,\ \ell \leq m,\ U_\ell \perp\!\!\!\perp U_m\},$$

where $\perp\!\!\!\perp$ denotes statistical independence.

The proof is in section A.1 of the mathematical Appendix.

Lemma 1 shows that several commonly used error dependence structures satisfy Assumption A2. A first example is provided by independent heteroskedastic errors. In this case:

$$\mathcal{J} = \{(\ell, m) \in \{1, ..., L\}^2, \ell < m\}, \quad \text{and} \quad J \equiv \#\mathcal{J} = \frac{L(L-1)}{2}.$$

If the data has a group structure, with r disjoint groups of size M_i $(i = 1, ..., r)$, and errors are independent between groups, then $\mathcal{J} = \{(\ell, m) \in M_i \times M_j, i \neq j, \ell < m\}$, and $J = \sum_{j=1}^{r-1} M_j \left(L - \sum_{i=1}^{j} M_i\right)$. The application to cognitive test scores, allowing for contemporaneous correlation in the errors, will offer an example of a block-independent structure.

In addition, Assumption A2 allows for temporal or spatial correlation patterns. For instance, if errors are MA(q) then $\mathcal{J} = \{(\ell, m) \in \{1, ..., L\}^2, \ell < m - q\}$, and $J = (L-q)(L-q-1)/2$. Likewise, spatial MA models may also satisfy the assumption, with J depending on the zeros of the matrix of spatial weights (e.g., Anselin, 2003).

In contrast, autoregressive (or spatial autoregressive) error structures do not satisfy Assumption A2, as errors are correlated at all lags and leads. However, the methods of this paper are applicable in this case also. To see how one might proceed, let us consider

a case where errors are ARMA(1,1). Then by taking quasi-differences $Y_\ell - \rho Y_{\ell-1}$, where ρ is the autoregressive parameter, we end up with MA(1) errors. Using the results below, ρ can then be obtained in the first estimation step (prewhitening), together with error moments.

2.2 Moment restrictions

We start by deriving the moment restrictions implied by Assumption A1. Let $p \in \{2,3,4\}$ and $(\ell_1, ..., \ell_p) \in \{1, ..., L\}^p$. Assumption A1 implies

$$\text{Cum}\left(Y_{\ell_1}, ..., Y_{\ell_p}\right) = \sum_{k=1}^{K} \left(\prod_{i=1}^{p} \lambda_{\ell_i,k} \right) \kappa_p\left(X_k\right) + \text{Cum}\left(U_{\ell_1}, ..., U_{\ell_p}\right), \tag{2}$$

where we write $\kappa_p(Z) = \text{Cum}(Z, ..., Z)$ (repeat Z p times) for univariate cumulants of order $p \geq 1$.

Moment restrictions (2) have a common multilinear structure which can be conveniently expressed in matrix form, as in ordinary Factor Analysis. Define the following L-by-L, symmetric, square matrices:

$$\begin{aligned}
\Sigma_Y &= \left[\text{Cum}\left(Y_i, Y_j\right)\right], \\
\Gamma_Y(\ell) &= \left[\text{Cum}\left(Y_i, Y_j, Y_\ell\right)\right], \quad \ell \in \{1, ..., L\}, \\
\Omega_Y(\ell, m) &= \left[\text{Cum}\left(Y_i, Y_j, Y_\ell, Y_m\right)\right], \quad \ell, m \in \{1, ..., L\},
\end{aligned}$$

with similar expressions for Σ_U, $\Gamma_U(\ell)$ or $\Omega_U(\ell, m)$.

Restrictions (2) imply that

$$\Sigma_Y = \Lambda\Lambda^{\mathrm{T}} + \Sigma_U, \tag{3}$$

$$\Gamma_Y(\ell) = \Lambda D_3 \,\text{diag}\left(\Lambda_\ell\right) \Lambda^{\mathrm{T}} + \Gamma_U(\ell), \tag{4}$$

$$\Omega_Y(\ell, m) = \Lambda D_4 \,\text{diag}\left(\Lambda_\ell \odot \Lambda_m\right) \Lambda^{\mathrm{T}} + \Omega_U(\ell, m), \tag{5}$$

where $\Lambda_\ell^{\mathrm{T}} \in \mathbb{R}^{K \times 1}$ is the ℓth row of Λ, D_3 (resp. D_4) is the diagonal matrix with cumulant $\kappa_3(X_k)$ (resp. $\kappa_4(X_k)$) in the kth entry of the diagonal, and $\odot$ is the Hadamard (element by element) matrix product.

Assumption A2 imposes additional restrictions. Combining the assumption with restrictions (5) yields:

$$\Omega_Y(\ell, m) = \Lambda D_4 \operatorname{diag}(\Lambda_\ell \odot \Lambda_m) \Lambda^{\mathrm{T}}, \quad \forall (\ell, m) \in \mathcal{J}.$$

For a symmetric matrix $A = [a_{ij}]$, we denote as vech the operator that stacks the elements of the upper triangular part of A, extracted horizontally from left to right: $\operatorname{vech}(A) = [a_{ij}]_{i \leq j}$. Applying the vech operator we obtain:

$$\operatorname{vech}(\Omega_Y(\ell, m)) = Q D_4 (\Lambda_\ell \odot \Lambda_m), \quad \forall (\ell, m) \in \mathcal{J},$$

where Q is the $\frac{L(L+1)}{2}$-by-K matrix which generic (i, j) row, $i \leq j$, is $(\lambda_{i1}\lambda_{j1}, ..., \lambda_{iK}\lambda_{jK})$, i.e.

$$Q \equiv \left[\operatorname{vech}\left(\boldsymbol{\lambda}_1 \boldsymbol{\lambda}_1^{\mathrm{T}}\right), ..., \operatorname{vech}\left(\boldsymbol{\lambda}_K \boldsymbol{\lambda}_K^{\mathrm{T}}\right)\right],$$

where $\boldsymbol{\lambda}_k$ denotes the kth column of Λ.

Next, construct the $\frac{L(L+1)}{2}$-by-J matrix Ω_Y by concatenating all vectors $\operatorname{vech}(\Omega_Y(\ell, m))$, $(\ell, m) \in \mathcal{J}$. Clearly:

$$\begin{aligned} \Omega_Y &\equiv [\omega_Y(\ell, m)]_{(\ell,m)\in\mathcal{J}} \\ &= [\operatorname{Cum}(Y_i, Y_j, Y_\ell, Y_m)]_{(i\leq j)\times(\ell,m)\in\mathcal{J}}. \end{aligned}$$

Matrix Ω_Y contains all fourth-order cumulants of measurements which are not contaminated by the presence of noise. Moreover, letting $Q_{\mathcal{J}}$ be the J-by-K matrix obtained by selecting rows $(i, j) \in \mathcal{J}$ from Q, we obtain:

$$\Omega_Y = Q D_4 {Q_{\mathcal{J}}}^{\mathrm{T}}. \tag{6}$$

We can similarly construct the following matrix of third-order cumulants:

$$\Gamma_Y = [\operatorname{Cum}(Y_i, Y_\ell, Y_m)]_{i\times(\ell,m)},$$

where the rows of Γ_Y are indexed by $i \in \{1, ..., L\}$ and the columns are indexed by $(\ell, m) \in \mathcal{J}$. Then,

$$\Gamma_Y = \Lambda D_3 {Q_{\mathcal{J}}}^{\mathrm{T}}. \tag{7}$$

In the next section, we take $\mathcal{J}$ as given and focus on the identification of factor loadings, using moment restrictions (3) to (7).

3 Identification results

In this section, we use the moment restrictions implied by the noisy ICA model to give sufficient conditions for the identification of factor loadings and error moments. We start with the number of factors, K.

3.1 Identification of the number of factors

The following theorem is an immediate consequence of (6) and (7).

Theorem 1 *The two following statements hold:*

i) Assume that all factor variables are kurtotic ($\kappa_4(X_k) \neq 0, \forall k$), and that matrix $Q_{\mathcal{J}}$ has rank K, which in particular implies $K \leq J$. Then matrix Ω_Y has rank K.

ii) Assume that all factors are skewed ($\kappa_3(X_k) \neq 0, \forall k$), and that both Λ and $Q_{\mathcal{J}}$ have rank K, which implies that $K \leq \min\{J, L\}$. Then Γ_Y has rank K.

Theorem 1 shows that matrices Ω_Y and Γ_Y allow to identify the number of common factors K. Notice that fourth-order cumulants can be used together with third-order cumulants. Define

$$\begin{aligned} \Omega_Y(j) &= [\mathrm{Cum}(Y_i, Y_j, Y_\ell, Y_m)], \ j \in \{1, ..., L\}, \text{ and} \\ \Phi_Y &= [\Gamma_Y, \Omega_Y(1), ..., \Omega_Y(L)]. \end{aligned}$$

Then, it is easily shown that, if factors are either skewed or kurtotic and Λ and $Q_{\mathcal{J}}$ have rank K, then matrix Φ_Y has rank K.

3.2 Identification of error moments

Applying operator vech to (3), (4) and (5) yields the following linear restrictions:

$$\begin{aligned}
\operatorname{vech}(\Sigma_Y) &= Q\mathbf{1}_K + \operatorname{vech}(\Sigma_U), \\
\operatorname{vech}(\Gamma_Y(\ell)) &= QD_3\Lambda_\ell + \operatorname{vech}(\Gamma_U(\ell)), \quad \forall \ell, \\
\operatorname{vech}(\Omega_Y(\ell,m)) &= QD_4(\Lambda_\ell \odot \Lambda_m) + \operatorname{vech}(\Omega_U(\ell,m)), \quad \forall(\ell,m),
\end{aligned}$$

where $\mathbf{1}_K$ is a K-dimensional vector of ones.

Let us begin by assuming that all factors are kurtotic, so that D_4 has no zero on its main diagonal. Theorem 1 shows that, if matrix $Q_{\mathcal{J}}$ has rank K, then $\operatorname{rank}(\Omega_Y) = K$. So one can choose an orthogonal basis of the null space of Ω_Y^{T}, and construct a $\frac{L(L+1)}{2}$-by-$\left(\frac{L(L+1)}{2} - K\right)$ orthogonal matrix B that satisfies: $\Omega_Y^{\mathrm{T}} B = 0$. Hence, as $Q_{\mathcal{J}} D_4$ has full column rank, it follows that $Q^{\mathrm{T}} B = 0$. So,

$$B^{\mathrm{T}} \operatorname{vech}(\Sigma_Y) = B^{\mathrm{T}} \operatorname{vech}(\Sigma_U), \tag{8}$$

$$B^{\mathrm{T}} \operatorname{vech}(\Gamma_Y(\ell)) = B^{\mathrm{T}} \operatorname{vech}(\Gamma_U(\ell)), \quad \forall \ell, \tag{9}$$

$$B^{\mathrm{T}} \operatorname{vech}(\Omega_Y(\ell,m)) = B^{\mathrm{T}} \operatorname{vech}(\Omega_U(\ell,m)), \quad \forall(\ell,m). \tag{10}$$

The following theorem shows that these linear restrictions identify error cumulants.

Theorem 2 *Assume that all factor variables have non zero excess kurtosis and that matrix $Q_{\mathcal{J}}$ has rank K. Then, second, third and fourth-order cumulants of error variables are uniquely defined by identifying restrictions (8), (9) and (10).*

The proof is in Section A.2 of the mathematical appendix.

Theorem 2 provides linear restrictions identifying error cumulants of order 2 to 4 irrespective of Λ and X. The theorem shows that high-order moments of the data, appearing in (4) and (5), contain information on error moments that is not contained in second-order moments of the data. Exploiting this information allows to increase the

number of common factors that can be identified in Factor Analysis, which only relies exclusively on second-order restrictions (3).

The following corollary is immediate.

Corollary 1 *Assume that the conditions of Theorem 2 are satisfied. Then the elements of* $\Lambda\Lambda^{\mathrm{T}}$ *are uniquely defined by restrictions (3) and (8).*

If $K \leq L$, the corollary shows that if the conditions of Theorem 2 hold, then Λ is identified up to right-multiplication by an orthogonal matrix. The last part of the identification proof, that we derive in the next section, is devoted to the identification of this rotation.

Corollary 1 can be of interest in its own right, if *ex-ante* restrictions are assumed on matrix Λ. Indeed, if these restrictions are sufficient to identify Λ from the knowledge of $\Lambda\Lambda^{\mathrm{T}}$, then the rest of the identification proof is unnecessary. This is for example the case if Λ is assumed to be lower triangular, as in the following linear panel data model:

$$y_{it} = p_{it} + u_{it}, \quad (i,t) \in \{1,...,N\} \times \{1,...,T\},$$

where p_{it} is a random walk: $p_{it} = p_{i,t-1} + \varepsilon_{it}$, with $p_{i0}, \varepsilon_{i1}, ..., \varepsilon_{iT}$ independent. The transitory shocks u_{it} can be e.g. MA(q), or the sum of an MA(q) and an iid component (e.g., measurement error).

We can proceed similarly if every factor is skewed. If both Λ and $Q_{\mathcal{J}}$ have full column rank K, Theorem 1 shows that $\Gamma_Y = \Lambda D_3 {Q_{\mathcal{J}}}^{\mathrm{T}}$ has rank K. Hence, there exists a L-by-$(L-K)$ orthogonal matrix C such that $\Gamma_Y^{\mathrm{T}} C = 0$. So, as D_3 has no zero on its diagonal, it must also be that $C^{\mathrm{T}}\Lambda = 0$.

The second, third and fourth-order cumulants of U_ℓ, for all $\ell \in \{1,...,L\}$, thus satisfy

the following linear restrictions:

$$C^{\mathrm{T}}\Sigma_Y = C^{\mathrm{T}}\Sigma_U, \tag{11}$$

$$C^{\mathrm{T}}\Gamma_Y(\ell) = C^{\mathrm{T}}\Gamma_U(\ell), \tag{12}$$

$$C^{\mathrm{T}}\Omega_Y(\ell, m) = C^{\mathrm{T}}\Omega_U(\ell, m). \tag{13}$$

Define, for all $\ell \in \{1, ..., L\}$, the sets

$$\mathcal{I}_\ell = \{m \in \{1, ..., L\},\ m < \ell \text{ or } (\ell, m) \in \mathcal{J}\},$$

with $I_\ell = \#\mathcal{I}_\ell$. Denote also $\Lambda_{\mathcal{I}_\ell}$ the I_ℓ-by-K matrix obtained by selecting rows $i \in \mathcal{I}_\ell$ from Λ. The following theorem gives conditions under which the system of linear restrictions (11), (12), and (13), has a unique solution.

Theorem 3 *Assume that every factor distribution is skewed, that $Q_{\mathcal{J}}$ has rank K, and that $\Lambda_{\mathcal{I}_\ell}$ has full column rank for all ℓ. Then second, third and fourth-order cumulants of error variables are identified from restrictions (11), (12), and (13).*

The proof is in Section A.3 of the mathematical appendix.

Theorem 3 implies that the number of factors is bounded by $\min\{I_\ell, \ell \in \{1, ..., L\}\}$. In the particular case of independent errors this yields $K \leq L-1$. Focusing on the model with $L = 2$ and $K = 1$, Geary (1942) has shown that identification holds, provided that the factor is skewed. Theorem 3 provides a generalization of this result to multi-factor models.

Lastly, the discussion in this subsection can be generalized to the case where every factor is either skewed or kurtotic ($\kappa_3(X_k)\,\kappa_4(X_k) \neq 0$). One needs only replace matrix Γ_Y by matrix $\Phi_Y = [\Gamma_Y, \Omega_Y(1), ..., \Omega_Y(L)]$, and compute C such that $\Phi_Y^{\mathrm{T}} C = 0$.

3.3 Identification of factor loadings

In this section we assume that the cumulants of order 2, 3 and 4 of error components are known, the previous section giving sufficient conditions for their identification. Second,

third and fourth-order restrictions (3), (4), (5) imply that matrix Λ satisfies, simultaneously,

$$\widetilde{\Sigma}_Y \equiv \Sigma_Y - \Sigma_U = \Lambda\Lambda^{\mathrm{T}}, \tag{14}$$

$$\widetilde{\Gamma}_Y(\ell) \equiv \Gamma_Y(\ell) - \Gamma_U(\ell) = \Lambda D_3 \operatorname{diag}(\Lambda_\ell)\Lambda^{\mathrm{T}}, \tag{15}$$

$$\widetilde{\Omega}_Y(\ell, m) \equiv \Omega_Y(\ell, m) - \Omega_U(\ell, m) = \Lambda D_4 \operatorname{diag}(\Lambda_\ell \odot \Lambda_m)\Lambda^{\mathrm{T}}. \tag{16}$$

Let us assume that $K \leq L$, and let P be an L-by-K matrix such that

$$P\widetilde{\Sigma}_Y P^{\mathrm{T}} = I_K. \tag{17}$$

Matrix P can easily be constructed from eigenvectors and eigenvalues of $\widetilde{\Sigma}_Y$. Left and right-multiplying (14), (15) and (16) by P and P^{T}, respectively, we obtain:

$$\begin{aligned} P\widetilde{\Gamma}_Y(\ell)P^{\mathrm{T}} &= VD_3 \operatorname{diag}(\Lambda_\ell)V^{\mathrm{T}}, \quad \ell \in \{1, ..., L\}, \\ P\widetilde{\Omega}_Y(\ell, m)P^{\mathrm{T}} &= VD_4 \operatorname{diag}(\Lambda_\ell \odot \Lambda_m)V^{\mathrm{T}}, \quad \ell \leq m, \end{aligned}$$

where $V = P\Lambda$ is orthonormal ($VV^{\mathrm{T}} = I_K$). Therefore, V solves a joint diagonalization problem. Theorem 4 below gives conditions for the solution to this problem to be unique.

Theorem 4 *Assume that error cumulants are known, and that matrix Λ has full column rank K, so in particular $K \leq L$.*

(i) If at most one factor variable has zero excess kurtosis, then factor loadings are identified from second and fourth-order moment restrictions (14) and (16).

(ii) If at most one factor variable has zero skewness, then factor loadings are identified from second and third-order moment restrictions (14) and (15).

(iii) If for any couple of factors indices (k, k'), $(\kappa_3(X_k), \kappa_3(X_{k'}), \kappa_4(X_k), \kappa_4(X_{k'})) \neq 0$, then factor loadings are identified from second, third and fourth-order moment restrictions (14), (15) and (16).

The proof is in Section A.4 the mathematical appendix.

Combining Theorems 2, 3 and 4 we obtain that (i) at most $K = \min\{J, L\}$ factors can be identified if all factors are kurtotic, and (ii) at most $K = \min\{I_\ell, \ell \in \{1, ..., L\}\}$ if all factors are either skewed or kurtotic. In the case where errors are independent one can thus identify up to $K = L$ factors in the first case and $K = L - 1$ in the second case. By comparison, the number of factors in FA models is bounded by $K = \left(2L + 1 - \sqrt{8L+1}\right)/2$. The general identification results hold provided that the errors are not too correlated. To give an example, if errors follow an MA(q) process indexed by the measurement indices $\ell = 1, ..., L$, then one can generically identify L common factors if $J = (L-q)(L-q-1)/2 \geq L$, that is if $q \leq (2L - 1 - \sqrt{8L+1})/2$.

We end this section by remarking that Lemma 2, together with the previous identification theorems, imply that overcomplete ICA models are identified if there exist sufficiently many restrictions on factor loadings. To see that, let us consider the model:

$$Y = \Lambda_1 X_1 + ... + \Lambda_S X_S + U,$$

where, for all $s \in \{1, ..., S\}$, X_s has $K_s \leq L$ elements, Λ_s is L-by-K_s, and all factors and errors are assumed mutually independent. Let us suppose that all factors are kurtotic, the argument being similar when factors are skewed. Theorems 2 and 4 show that one can generically identify up to $K_1 = J_1$ factors X_1, where J_1 is the number of components of $\Lambda_2 X_2 + ... + \Lambda_S X_S + U$ that are mutually independent. As an example, $K_1 = L - 1$ factors X_1 are identified, if the first row of all matrices $\Lambda_2, ..., \Lambda_S$ is identically zero. Applying this procedure sequentially shows identification in the case where $S = L - 1$, and for all $s \in \{1, ..., S\}$ $K_s = L - s$, and the first $s - 1$ rows of Λ_s are zero. This corresponds to a block-triangular structure where the first $L - 1$ factors are common to all measurements, the next $L - 2$ factors are specific to $Y_2, ..., Y_L$, and so on. In this model there are $K = L(L-1)/2$ factors, and $L(L-1)^2/6$ restrictions on the $L^2(L-1)/2$ factor loadings.

4 Estimation

We start by discussing the issue of estimating the number of factors, and the factor loadings. Then, we provide the asymptotic theory of the JADE estimator, and discuss how to perform inference for JADE and quasi-JADE in practice.

4.1 Estimating the number of factors K

All factors are kurtotic. Assuming that $Q_{\mathcal{J}}$ has full column rank and that factor variables show excess kurtosis, then matrix Ω_Y has rank $K \leq J$ (see Theorem 1). We use the sequential testing procedure developed by Robin and Smith (2000) to estimate the rank of Ω_Y (see Appendix D for a description of the test).

Monte Carlo simulations show that the rank test, applied to matrix Ω_Y alone, suffers from substantial size distortions (see the simulations in the next section). Assuming $K \leq L$, the factor structure provides additional rank conditions that can be used to improve the test's properties. We propose the following refinement.

Consider matrices $\Omega_Y(\ell, m)$ with $(\ell, m) \in \mathcal{J}$. They satisfy the restrictions:

$$\Omega_Y(\ell, m) = \Lambda D_4 \operatorname{diag}(\Lambda_\ell \odot \Lambda_m) \Lambda^{\mathrm{T}}.$$

Let $w = (w_{\ell m}, (\ell, m) \in \mathcal{J})$ be a vector of J positive weights. Then,

$$\Omega_{Y,w} \equiv \sum_{(\ell,m)\in\mathcal{J}} w_{\ell,m} \Omega_Y(\ell, m) = \Lambda D_4 \operatorname{diag}\left({Q_{\mathcal{J}}}^{\mathrm{T}} w\right) \Lambda^{\mathrm{T}}. \tag{18}$$

As no column of $Q_{\mathcal{J}}$ is identically zero, matrix $\Omega_{Y,w}$ has rank K for almost all w.

It seems natural to weight cumulant matrices more if they are more precise. We therefore suggest to choose $w_{\ell,m}$ equal to the inverse of the simple average of the asymptotic variances of the components of an empirical analog $\widehat{\Omega}_Y(\ell, m)$ of $\Omega_Y(\ell, m)$. These variances can be computed by standard bootstrap.

All factors are skewed (or either skewed or kurtotic). Assuming that Λ and $Q_{\mathcal{J}}$ have full column rank and that factor variables have non zero skewness, then matrix Γ_Y has rank $K \leq \min(I_\ell, \ell \in \{1, ..., L\})$. One can thus apply the rank test to any analog estimator $\widehat{\Gamma}_Y$.

Assuming that each factor distribution is either skewed or kurtotic, matrix

$$\Phi_Y = [\Gamma_Y, \Omega_Y(1), ..., \Omega_Y(L)]$$

has rank K. One can thus test the rank of any consistent analog estimator $\widehat{\Phi}_Y$. Alternatively, in the same spirit as in the previous paragraph, remark that, under the assumption that all factors are skewed or kurtotic, all matrices

$$\Phi_{Y,w} = \Gamma_Y + \sum_{j=1}^{L} w_j \Omega_Y(j) = \Lambda \left[D_3 + D_4 \operatorname{diag}\left(\Lambda^{\mathrm{T}} w\right) \right] {Q_{\mathcal{J}}}^{\mathrm{T}} \tag{19}$$

have rank K, for almost all weights $w = (w_1, ..., w_L)^{\mathrm{T}} \in \mathbb{R}^L$. Matrices $\Phi_{Y,w}$ can therefore be used to estimate the number of factors K. We suggest to set w_j equal to the average of the variances of the components of $\widehat{\Gamma}_Y$ divided by the average of the variances of the components of $\widehat{\Omega}_Y(j)$.

4.2 Estimation of factor loadings

The two steps of the estimation algorithm are as follows.

Prewhitening. In the first step error moments are estimated. In the case where all factors are kurtotic one may apply the following procedure:

1. Construct matrix $\Omega_Y = [\mathrm{Cum}(Y_i, Y_j, Y_\ell, Y_m)]$, where rows are indexed by couples (i, j), $i \leq j$, and columns are indexed by couples $(\ell, m) \in \mathcal{J}$.

2. Assuming that $\mathrm{rank}(\Omega_Y) = K$, find the null space of Ω_Y^{T}, i.e. compute an orthogonal, $\frac{L(L+1)}{2}$-by-$\left(\frac{L(L+1)}{2} - K\right)$ matrix B such that $\Omega_Y^{\mathrm{T}} B = 0$. A Singular Value Decomposition (SVD) can be used for this purpose.

3. Solve for the non-zero elements of Σ_U in the linear system (8). Proceed in the same way for third-order and fourth-order error cumulant matrices $\Gamma_U(\ell)$ and $\Omega_U(\ell, m)$.

Alternatively, if all factors are skewed, or either skewed or kurtotic, one can follow a similar procedure, basing the estimation on matrix Γ_Y or matrix Φ_Y, respectively.

In the algorithm, Step 3 can be performed by Least Squares. However, doing so does not necessarily deliver a positive-definite matrix $\Sigma_Y - \Sigma_U$. This is why it seems preferable to combine the linear restrictions (8) with the covariance restrictions (3), and perform a factor analysis of Σ_Y with linearly constrained error variances and covariances.

In practice, we simultaneously solve for the lower triangular matrices W (L-by-K) and Z (L-by-L) such that restrictions

$$\begin{aligned} \Sigma_Y &= WW^{\mathrm{T}} + ZZ^{\mathrm{T}}, \\ B^{\mathrm{T}} \operatorname{vech}(\Sigma_Y) &= B^{\mathrm{T}} \operatorname{vech}\left(ZZ^{\mathrm{T}}\right), \\ \left[ZZ^{\mathrm{T}}\right]_{(\ell,m)} &= 0, \quad \forall (\ell, m) \in \mathcal{J}, \end{aligned}$$

approximately hold in a L^2 sense. This is a quadratic problem that can be solved using standard optimization routines.

Remark that, if there are sufficently many restrictions on Λ for it to be identified, then one can estimate Λ together with Σ_U directly from this system. The source separation step below then becomes unnecessary, see the discussion following Corollary 1.

Because one can base the restrictions on error moments either on Ω_Y or Γ_Y, we suggest to weight both sets of moment restrictions by the average of the variances of the components of an estimate $\widehat{\Sigma}_Y$ divided by the average of the variances of the components of $\widehat{\Omega}_Y$ (resp. $\widehat{\Gamma}_Y$). See Cragg (1997) for a related strategy based on the moments of the standard normal distribution.

Lastly, once errors cumulants Σ_U, $\Gamma_U(\ell)$ and $\Omega_U(\ell, m)$ have been estimated, the data cumulant matrices are whitened as in equations (14), (15) and (16).

Source separation. Given whitened cumulant matrices $\widetilde{\Sigma}_Y$, $\widetilde{\Gamma}_Y(\ell)$ and $\widetilde{\Omega}_Y(\ell, m)$, we compute V as the K-by-K matrix of common orthonormal eigenvectors ($VV^{\mathrm{T}} = I_K$) of matrices $P\widetilde{\Gamma}_Y(\ell)\,P^{\mathrm{T}}$ and $P\widetilde{\Omega}_Y(\ell, m)\,P^{\mathrm{T}}$, where P satisfies equation (17). For example, one can choose $P = W^{-}$ (the Moore-Penrose generalized inverse of W), where W has been estimated in the prewhitening step. Then factor loadings are obtained as $\Lambda = PV$.

In practice, we replace theoretical moments by empirical ones and use Cardoso and Souloumiac's (1993) Joint Approximate Diagonalization algorithm (JADE). This algorithm provides a fast way of minimizing with respect to an orthonormal matrix V the sum of squares of off-diagonal elements of matrices $V^{\mathrm{T}}P\widetilde{\Gamma}_Y(\ell)\,P^{\mathrm{T}}V$ and $V^{\mathrm{T}}P\widetilde{\Omega}_Y(\ell, m)\,P^{\mathrm{T}}V$. As before one may weight the third and fourth-order cumulants differently. We suggest to use the same weights as in the prewhitening step.

The JADE algorithm is described in Appendix B. We call the resulting algorithm quasi-JADE, to emphasize the two-step nature of our procedure. It is only marginally more complicated to implement than JADE and almost as fast. However, unlike JADE, it is robust to the presence of (possibly correlated) noise.

Lastly, once factor loadings have been estimated, one can obtain the third and fourth-order cumulants of factor variables from the linear restrictions (6) and (7).

4.3 Inference

As far as we know, there is no derivation of the asymptotic properties of JADE in the ICA literature. This section aims at filling this gap. At the end of the section, we discuss how to perform inference for the JADE and quasi-JADE estimates in practice.

To proceed, let $\widehat{A}_1, ..., \widehat{A}_S$ be root-N consistent and asymptotically normal estimators of S symmetric K-by-K matrices $A_1, ..., A_S$. Construct $\widehat{A} = \left[\widehat{A}_1, ..., \widehat{A}_S\right]$ and $A = [A_1, ..., A_S]$ by concatenation. Let $\mathbb{V}_A$ be the asymptotic variance of $N^{\frac{1}{2}}\,\mathrm{vec}(\widehat{A})$. The

JADE estimator is

$$\widehat{V} = \arg\min_{V \in \mathcal{O}_K} \sum_{s=1}^{S} \text{off}(V^{\mathrm{T}} \widehat{A}_s V),$$

where $\text{off}(M) = \sum_{i \neq s} m_{is}^2$ for a matrix $M = [m_{is}]$, and $\mathcal{O}_K$ is the set of orthonormal K-by-K matrices.

Assume that there exists $V \in \mathcal{O}_K$ such that, for all $s = 1, ..., S$, $V^{\mathrm{T}} A_s V = D_s$, where D_s is the diagonal matrix with diagonal elements $d_{s1}, ..., d_{sK}$. Define the K-by-K matrices:

$$R_s = \left[\frac{(d_{sk} - d_{sm})}{\sum_{s'=1}^{S} (d_{s'k} - d_{s'm})^2} \right]_{k,m=1,...,K},$$

and $r_s = \text{vec}(R_s)$. Lastly, let W be the following K^2-by-SK^2 matrix:

$$W = \left[\text{diag}\left(r_1\right), ..., \text{diag}\left(r_S\right) \right].$$

We show the following result in Appendix C.

Theorem 5 *Assume that* $\sum_{s=1}^{S} (d_{sk} - d_{sm})^2 \neq 0$ *for all* $k \neq m$. *Then*

$$N^{\frac{1}{2}} \left(\text{vec}(\widehat{V}) - \text{vec}(V) \right) \rightarrow \mathcal{N}\left(0, \mathbb{V}_V\right) \quad \textit{(weakly)},$$

where:

$$\mathbb{V}_V = (I_K \otimes V) W (I_S \otimes V^{\mathrm{T}} \otimes V^{\mathrm{T}}) \mathbb{V}_A (I_S \otimes V \otimes V) W^{\mathrm{T}} (I_K \otimes V^{\mathrm{T}}). \tag{20}$$

Let us consider the particular case of $S = 1$. In this case, (20) yields the well-known expression for the variance-covariance matrix of the eigenvectors of a symmetric matrix (e.g., Anderson, 1963). The diagonal coefficients of matrix W are equal to $1/(d_{1k} - d_{1m})$, for $k \neq m$. The variance of eigenvectors thus increases when two eigenvalues of A_1 get close to each other.

In the general case of more than one matrix ($S > 1$), precise estimation requires $\sum_s (d_{sk} - d_{sm})^2$ to be away from zero, for all indices (k, m). Cardoso (1999) already noted that joint diagonalization algorithms seemed less sensitive to the presence of multiple roots than usual diagonalization techniques (see also the asymptotic distribution

of estimators of Common Principal Components derived by Flury, 1986). Theorem 5 permits to better understand the conditions granting a good precision.

In the quasi-JADE algorithm using fourth-order moments, indices are $s = (\ell, m)$, and matrices D_s are: $D_4 \operatorname{diag}(\Lambda_\ell \odot \Lambda_m)$. If there exist k, k' such that $d_{sk} = d_{sk'}$ for all s, it must be that

$$\lambda_{\ell k}\lambda_{mk}\kappa_4(X_k) = \lambda_{\ell k'}\lambda_{mk'}\kappa_4(X_{k'}), \forall(\ell, m).$$

This cannot happen if at most one factor has zero excess kurtosis and the columns of Λ are not proportional to each other.

This result is not surprising, as the variance of eigenvector estimators blows up when the model is not identified. Non identification arises in PCA when the variance of the vector of measurements has multiple eigenvalues (there are then obviously many possible choices for a basis of the corresponding eigenspace). In ICA this happens when two columns of the matrix of factor loadings are proportional or when factor distributions lack skewness and/or excess kurtosis. We shall produce Monte-Carlo simulations to illustrate this point.

Lastly, the asymptotic result for JADE given in Theorem 5 can be generalized to quasi-JADE, at the cost of introducing extra notation. As a result, the algorithm yields root-N consistent and asymptotically normal estimates of factor loadings and error moments, under the conditions of Theorem 2 (or Theorem 3, if using third-order moments) and Theorem 4. However, this generalization is not of direct interest to our purpose, as illustrated by the next remark.

Practical remark. In practice, we do *not* recommend to use formula (20) to compute standard errors. Instead, we suggest to compute standard errors or confidence intervals by standard bootstrap (maybe with appropriate recentering for finite sample improvements). The reason is that (20) involves variances of third and/or fourth-order moments of the

Table 1: Empirical cumulants of the standard log-normal

N	500	1000	5000	10000
κ_3	4.49 (2.20)	4.87 (2.47)	5.66 (2.54)	5.83 (3.17)
κ_4	35.9 (.88)	44.5 (.93)	72.9 (.72)	79.8 (.93)
κ_6	4,825 (.36)	8,698 (.35)	44,492 (.21)	55,505 (.28)
κ_8	856,819 (.22)	2,642,849 (.20)	59,108,559 (.12)	80,815,329 (.16)

Note: Empirical skewness, excess kurtosis, 6th and 8th-order cumulants of a log-normal random variable. t-statistics in parentheses. Estimates from 1000 independent draws, for each sample size N.

data, i.e. sixth and eighth-order moments. These are difficult to estimate precisely (see Table 1 for an example with log-normal variables). In our simulation experiments, we obtained extremely imprecise estimates of matrix $\mathbb{V}_A$, even with very large samples (more than $10,000$ observations). In contrast, bootstrapping provided good approximations of the true variance-covariance matrix of the JADE estimator.

5 Monte-Carlo simulations

In this section, we study the finite-sample properties of our estimator with numerical simulations. We first consider the estimation of Λ given the true value of K, the number of factors. Then, we turn to the estimation of K.

5.1 Estimation of factor loadings

Table 2 displays means and standard deviations of the Monte Carlo distributions of factor loadings estimates obtained from 1000 simulations of samples of various sizes generated by standardized log-normal factors, standard normal errors and Λ equal to

$$\Lambda_1 \equiv \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}.$$

We only report the estimates of the first column of Λ and the variance of the first error, the other estimates being qualitatively similar. Monte Carlo standard deviations of estimates

Table 2: Quasi-JADE for various sample sizes

N	500	1000	5000	10000
λ_{11}	2.03 (.28)	2.03 (.17)	2.01 (.09)	2.01 (.06)
λ_{21}	.95 (.23)	.99 (.14)	1.00 (.07)	1.00 (.05)
λ_{31}	.95 (.23)	.99 (.15)	.99 (.07)	1.00 (.05)
$\text{Var}(U_1)$	.77 (.59)	.87 (.43)	.96 (.20)	.98 (.16)

Note: log-normal factors, standard normal errors, $\Lambda = \Lambda_1$.

are given between brackets. Estimation is based on all moments of order 2, 3 and 4 of the data and uses the restrictions of Theorem 2. The error moments are estimated by least squares, based on restrictions (8)-(10) and (11)-(13).

Table 2 shows that finite sample biases are small and rapidly decrease as N increases. By comparison, small sample biases are much larger and convergence is much slower for empirical cumulants. The striking contrast between Tables 2 and 1 suggests that our algorithm does a good job at extracting the relevant information from high-order moments of the data, while being relatively immune to the imprecision of their estimation in finite samples.

We then study the robustness of the JADE and quasi-JADE algorithms to noise (see Table 3). We run the simulations with normal errors, log-normal factors, a sample size of $N = 1000$ and $\Lambda = \Lambda_1$. The standard deviation of errors can take four values: 0.1, 0.5, 1 and 2. The performance of quasi-JADE deteriorates as the signal-to-noise ratio decreases. However, biases remain limited even for rather large error variances. By comparison, JADE produces large finite sample biases.

Next, we investigate the sensitivity of our algorithm to factor Gaussianity. The sample size is $N = 1000$. Errors are standard normal variables. We simulate symmetric, kurtotic factors as mixtures of two independent normals. Table 4 summarizes Monte Carlo distributions for kurtosis values in $\frac{1}{2}$, 2, 5, 10 and 100. In the first column of Table 4, we also report results for the case of uniformly distributed factors. The uniform dis-

Table 3: Robustness to noise

JADE				
$\text{Var}(U_\ell)$	.01	.25	1	4
λ_{11}	2.00 (.07)	2.11 (.08)	2.36 (.12)	2.81 (.46)
λ_{21}	1.00 (.11)	1.00 (.12)	.95 (.24)	.72 (.86)
λ_{31}	1.00 (.11)	1.03 (.14)	1.08 (.22)	1.05 (.77)

Quasi-JADE				
$\text{Var}(U_\ell)$	.01	.25	1	4
λ_{11}	1.98 (.12)	2.01(.13)	2.03 (.17)	2.02 (.44)
λ_{21}	1.00 (.15)	.99 (.12)	.99 (.14)	.95 (.31)
λ_{31}	1.00 (.16)	.99 (.13)	.99 (.15)	.95 (.32)
$\text{Var}(U_1)$	.04 (.11)	.18 (.22)	.87 (.43)	3.77 (.98)

Note: log-normal factors, standard normal errors, $\Lambda = \Lambda_1$, $N = 1000$.

Table 4: Near-Gaussianity biases

κ_4	-6/5	1/2	1	5	10	100	≈110
	(uniform)			(normal mixtures)			(log-normal)
λ_{11}	1.94 (.48)	1.66 (.78)	1.76 (.74)	2.03 (.33)	2.01 (.26)	2.01 (.19)	2.03 (20)
λ_{21}	.91 (.48)	.97 (.71)	.94 (.63)	.97 (.30)	.98 (.21)	.99 (.16)	.98 (.15)
λ_{31}	.92 (.48)	1.00 (.69)	.96 (.65)	.97 (.29)	.97 (.21)	.98 (.17)	.98 (.16)
$\text{Var}(U_1)$	.71 (.65)	.92 (.84)	.76 (.79)	.77 (.63)	.88 (.53)	.92 (.40)	.86 (.44)

Note: factors are normal mixtures, standard normal errors, $\Lambda = \Lambda_1$, $N = 1000$.

tribution is platykurtic, with $\kappa_4 = -6/5$. The last column shows results for log-normal factors, with excess kurtosis equal to $e^4 + 2e^3 + 3e^2 - 6 \approx 110$. Overall, we find that the impact of kurtosis on the performance of the algorithm is far from negligible. The closer the excess kurtosis is to zero, the greater the estimator's bias and the lower its precision.

We now set $K < L$ and compare quasi-JADE based on second, third and fourth-order moments (using the restrictions of Theorem 2) to quasi-JADE based on second and third-order moments only (using the restrictions of Theorem 3), which yields consistent estimates when all factors are skewed. Table 5 reports simulations with log-normal factors, standard normal errors with variance 1, and matrix Λ is equal to

$$\Lambda_2 \equiv \begin{pmatrix} 2 & 2 \\ 2 & 1 \\ 1 & 2 \end{pmatrix}. \tag{21}$$

Table 5 shows that the standard deviations of factor loadings estimates *increase* when adding fourth-order moments. This is because, in our design, the (finite-sample) imprecision of kurtosis estimates dominates the (asymptotic) efficiency gains of using more moments. This illustrative table suggests that an algorithm based on third-order moments only, and relying on orthogonality up to the third order, is likely to do well in practice, provided that there is enough factor skewness.

Then, we investigate the finite-sample performance of our algorithm when the number of measurements and the number of factors increase. Table 6 illustrates the cases $L = K = 5$ and $L = K = 10$, respectively. In both cases, Λ has entries equal to 2 everywhere on the diagonal, and equal to one everywhere else. These simulations show that the performance of our algorithm is only moderately damped by the number of factors/measurements. We view this as quite remarkable a result as a hundred of factor loadings is certainly a significant number of parameters to estimate given that no explanatory variable is observed. In comparison, standard gradient algorithms for non-linear method-of-moments estimators turn out to be impractical for L as low as five.

Table 5: Efficiency gains from using fourth order moments

N	500	500	1000	1000	5000	5000
Cumulants	2,3,4	2,3	2,3,4	2,3	2,3,4	2,3
λ_{11}	1.95 (.28)	1.93 (.32)	1.98 (.19)	1.97 (.24)	2.00 (.08)	2.00 (.08)
λ_{21}	1.96 (.30)	1.91 (.37)	1.99 (.16)	1.96 (.23)	1.00 (.09)	2.00 (.05)
λ_{31}	.97 (.23)	.98 (.25)	.98 (.17)	.98 (.20)	1.00 (.08)	1.00 (.08)
$\mathrm{Var}(U_1)$	.98 (.21)	1.01 (.16)	.98 (.15)	1.00 (.13)	.97 (.09)	1.00 (.06)

Note: log-normal factors, standard normal errors, $\Lambda = \Lambda_2$.

Table 6: Increasing the number of factors and measurements

	$L = K = 5$			$L = K = 10$		
N	500	1000	5000	500	1000	5000
λ_{11}	2.06 (.41)	2.03 (.28)	2.01 (.13)	1.85 (.72)	1.97 (.56)	2.00 (.27)
λ_{21}	.95 (.35)	.98 (.25)	.99 (.12)	.89 (.52)	.90 (.43)	.98 (.22)
λ_{31}	.95 (.34)	.98 (.24)	1.00 (.12)	.88 (.53)	.90 (.45)	.98 (.23)
λ_{41}	.95 (.35)	.98 (.24)	.99 (.11)	.88 (.53)	.92 (.43)	.98 (.22)
λ_{51}	.95 (.34)	.98 (.24)	.99 (.12)	.88 (.53)	.90 (.43)	.98 (.22)
λ_{61}				.88 (.54)	.91 (.43)	.98 (.22)
λ_{71}				.89 (.53)	.90 (.44)	.98 (.22)
λ_{81}				.88 (.52)	.90 (.44)	.98 (.23)
λ_{91}				.87 (.53)	.91 (.44)	.98 (.23)
$\lambda_{10,1}$				.88 (.52)	.89 (.44)	.98 (.22)
$\mathrm{Var}(U_1)$	.58 (.56)	.81 (.44)	.95 (.20)	.40 (.55)	.49 (.53)	.88 (.28)

Note: log-normal factors, standard normal errors.

Computing time becomes prohibitive and algorithms fail to converge in many cases.

Next, we consider a case with correlated errors. Table 7 displays means and standard deviations of the Monte Carlo distributions of factor loadings estimates obtained from 1000 simulations of samples of size $N = 1000$ generated by standardized log-normal factors, standard normal errors and Λ equal to

$$\Lambda_3 \equiv \begin{pmatrix} 2 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 \\ 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 2 \end{pmatrix}.$$

We only allow errors U_3 and U_4 to have non-zero correlation ρ. In the simulation we let ρ vary between 0 and .9. When ρ increases, the performances of the algorithms

Table 7: JADE and Quasi-JADE for various correlation parameter ρ

JADE				
ρ	0	.2	.5	.9
λ_{41}	.89 (.42)	.87 (.44)	.84 (.45)	.75 (.50)
λ_{44}	2.25 (.29)	2.22 (.32)	2.23 (.28)	2.20 (.24)
Quasi-JADE, independent errors				
ρ	0	.2	.5	.9
λ_{41}	.98 (.17)	.95 (.19)	.90 (.21)	.86 (.28)
λ_{44}	2.05 (.20)	2.08 (.19)	2.15 (.16)	2.15 (.23)
$\mathrm{Var}(U_4)$	.81 (.42)	.66 (.40)	.37 (.33)	.12 (.20)
Quasi-JADE, correlation allowed between U_3 and U_4				
ρ	0	.2	.5	.9
λ_{41}	.98 (.19)	.99 (.17)	.99 (.17)	.98 (.16)
λ_{44}	2.03 (.21)	2.03 (.23)	2.03 (.22)	2.05 (.22)
$\mathrm{Var}(U_4)$	.84 (.54)	.82 (.55)	.81 (.55)	.79 (.56)
$\mathrm{Cov}(U_3, U_4)$	-.002 (.22)	.20 (.23)	.49 (.24)	.88 (.25)

Note: log-normal factors, standard normal errors, $\Lambda = \Lambda_3$, $N = 1000$.

assuming no or independent errors deteriorate. By comparison, quasi-JADE, with the right structure of error dependences, shows slightly larger standard errors when $\rho = 0$, which is consistent with the fact that it uses less moment conditions to estimate error moments. When ρ increases, it turns out to be remarkably robust. In all cases the performance of noise-free JADE is much worse.

In our last experiment, we simulated an overcomplete ICA model with $L = 4$ and $K = 6$, with four restrictions on Λ:

$$\Lambda_4 \equiv \begin{pmatrix} 1 & 2 & 1 & 1 & 1 & 0 \\ 1 & 1 & 2 & 1 & 1 & 0 \\ 0 & 1 & 1 & 2 & 1 & 1 \\ 0 & 1 & 1 & 1 & 2 & 1 \end{pmatrix}.$$

We simulated the model 1000 times, with sample size $N = 1000$, standardized log-normal factors and standard normal errors. We obtained the following estimates (standard errors

Table 8: Size of the rank tests based on Ω_Y, $\Omega_{Y,w}$ and Γ_Y for increasing kurtosis

$\kappa_4(\rho)$	-6/5	1/2	1	5	10	100	110	110	110
	(uniform)	(normal mixtures)					(log-normal)		
	Test based on Ω_Y							$\Omega_{Y,w}$	Γ_Y
α =.10	.90	.73	.82	.87	.85	.62	.56	.87	.90
α =.20	.79	.57	.67	.74	.69	.43	.34	.71	.79
α =.50	.47	.24	.32	.40	.35	.11	.08	.32	.48
α =.90	.10	.02	.04	.06	.04	.00	.00	.01	.07

Note: factors are uniform, normal mixtures or log-normal, errors are Gaussian, $\Lambda = \Lambda_2$, $N = 1000$.

in parentheses):

$$\widehat{\Lambda}_4 \equiv \begin{pmatrix} 1.01 & 1.93 & .93 & .99 & 1.01 & \\ (.38) & (.45) & (.28) & (.31) & (.33) & 0 \\ 1.00 & .93 & 1.94 & .98 & 1.00 & \\ (.38) & (.31) & (.40) & (.30) & (.31) & 0 \\ & .97 & .98 & 1.97 & .97 & .94 \\ 0 & (.33) & (.32) & (.41) & (.29) & (.39) \\ & .96 & .98 & .95 & 1.98 & .95 \\ 0 & (.33) & (.30) & (.29) & (.41) & (.39) \end{pmatrix}.$$

Finite sample biases are somewhat larger than in the case $K \leq L$. Nevertheless, this result shows that quasi-JADE can be used to estimate restricted overcomplete ICA models, even when the sample size is only moderately large.

5.2 Estimation of the number of factors

We here report a Monte-Carlo study of the rank tests detailed in 4.1. We first compute the empirical size of the test based on matrix Ω_Y for various values of factor kurtosis. The simulation design is the same as for the results reported in Table 4. The true value of Λ is Λ_2 given by (21), and we test $K = 2$ against $K = 3$.

The first seven columns of Table 8 show substantial size distortion. This especially happens when excess kurtosis is low (in absolute value) – that is, when fourth-order cumulants contain very little information on the factor structure – or large – that is, when fourth-order moments are imprecisely estimated. However, for intermediate values

Table 9: Power of the improved rank test based on $\Omega_{Y,w}$

$\kappa_4(\rho)$	-6/5	1/2	1	5	10	100
	(uniform)		(normal mixtures)			
α =.10	.99	.81	.81	1.00	1.00	.89
α =.20	.99	.63	.66	1.00	1.00	.80
α =.50	.96	.26	.29	.98	.99	.56
α =.90	.83	.02	.04	.72	.77	.12

Note: factors are normal mixtures, standard normal errors, $\Lambda = \Lambda_2$, $N = 1000$.

of excess kurtosis the risk of underestimating the number of factors exists but remains limited.

In Section 4.1, we proposed to improve the size properties of the rank test by considering a weighted average of cumulant matrices $\Omega_Y(\ell, m)$ – i.e. $\Omega_{Y,w}$ in equation (18) – instead of Ω_Y. Column 8 in Table 8 shows that weighting scheme definitely improves the size of the test of $K = 2$ against $K = 3$. However, the rank test still under-rejects noticeably, in particular when the theoretical probability of rejection is low. Lastly, the last column refers to matrix Γ_Y (third-order cumulants). Third-order moments being more precisely estimated, the empirical size of the rank test based on Γ_Y is close to the nominal size (third column).

This confirms that applying the characteristic root test to matrices of high-order cumulants should be done with some caution when they are too imprecisely estimated. However, the results in Table 8 show that, when skewness and excess kurtosis are not too large, the size properties of the rank test based on third and fourth-order cumulant matrices are satisfactory.

We end this section by a study of the power of the rank test based on $\Omega_{Y,w}$. Table 9 displays empirical power computations for various levels of kurtosis. The true value of Λ is Λ_1 and we test $K = 2$ against $K = 3$. For low levels (α less than 10%) the power of the test is good even if factors are strongly leptokurtic. For intermediate values of excess

kurtosis, the power is good whatever the level.

6 Factor analysis on test scores data

In this section we apply our methodology to British data on cognitive test scores.

6.1 The data

Table 10: Descriptive statistics

	Mean	Variance	Skewness	Ex. kurtosis	N
Math (7)	53.4	589	.040	-.77	7816
Reading (7)	79.9	480	-1.17	.42	7816
Math (11)	44.4	657	.21	-1.07	7816
Reading (11)	47.8	302	.087	-.47	7816
Verbal (11)	58.2	509	-.23	-.92	7816
Math (16)	42.6	498	.46	-.68	7816
Reading (16)	75.0	329	-.89	.34	7816
Years left education	17.5	5.56	1.70	2.64	5653
Log monthly wage (2000)	4.51	.622	-.74	2.68	4012
Log hourly wage (2000)	.945	.308	-.59	8.65	3982
Female dummy	.492	.250	.03	-2.00	7816

Note: sample taken from the NCDS data, years 1965, 1969, 1974 and 2000.

The NCDS is a longitudinal survey of a British birth cohort born in the same week of 1958. We use the following waves: 1965 (age 7), 1969 (age 11), 1974 (age 16), and 2000 (age 42). To select the sample we consider all individuals for whom we have information on test scores for the first three waves. There are seven available test measures: mathematics and reading at age 7, 11 and 16, and a verbal test at age 11 only. We also use the age at the time of leaving school and the logarithms of monthly and hourly wages measured at age 42.

Table 10 shows the first moments of the variables of interest. For interpretability, we have rescaled the test score measures so that they range between 0 and 100 points. We remark that most test scores present some skewness, either right or left, and negative

excess kurtosis. In Table 11 we show the correlations between the seven test score measures and the years of education and log wage variables. We see that these correlations are all positive. Moreover, more recent scores and scores in mathematics appear more strongly correlated with later outcomes. Lastly, girls do slightly better in reading/verbal, and slightly worse in mathematics, than boys.

Table 11: Correlations between test scores and education, log wage and gender

	Years education	Log monthly wage	Log hourly wage	Female
Math (7)	.26	.20	.19	-.06
Reading (7)	.29	.10	.13	.12
Math (11)	.46	.25	.27	-.03
Reading (11)	.46	.24	.26	-.01
Verbal (11)	.39	.15	.19	.11
Math (16)	.53	.30	.31	-.11
Reading (16)	.42	.25	.26	-.03

6.2 Results

We analyze the data with an independent factor model. In the present context, it is natural to allow for errors, the distributions of which are *a priori* different for each test measure. Moreover, as the exams were given on the same day (in the three interviews) it makes sense to allow for contemporaneous correlation between test scores. For instance, it could be that a child had a "bad day" and performed badly in all the tests. For this reason, we allow for correlation between the errors in the reading and mathematics scores at age 7 and age 16, and between the reading, mathematics and verbal scores at age 11.

Our approach requires that the data be sufficiently non-normal. The moments reported in Table 10 show that there is some non-normal skewness and kurtosis in the marginal distributions of the score variables. In order to check the extent of non-normality in the joint distribution of test scores we performed the tests outlined in 4.1. The results of the rank tests based on third and fourth-order moments imply that we can reject

the restrictions imposed by a 5-factor model, while a 6-factor model cannot be rejected. Nevertheless, the estimates based on four-factor and five-factor models turned out to be very imprecise. So, in the following we present the results for one to three factors. In the estimation we use second, third and fourth-order moments jointly. To account for the fact that lower-order moments are better estimated, we weight the cumulant matrices as explained in 4.2. Relative to second-order moments, third-order moments are thus weighted by a factor .178, and fourth-order ones by .091.

Table 12 shows the estimation results. The first three columns show the factor loadings estimates that correspond to each of the seven test score measures. The last seven columns give the estimates of the variance-covariance matrix of error variables. The last two rows give the skewness and excess kurtosis of the factors. Lastly, bootstrap standard errors are given in parentheses (100 iterations). To interpret the factor variables, we give in Table 13 the correlation between the linear projection of test scores on factor loadings ($\widehat{X} \equiv \Lambda^{-} Y$), and the variables of interest. We interpret $\widehat{X}$ as an estimate of the vector of factor variables, though a more correct approach would consist in filtering X out using the independence assumptions (see the Conclusion).

Table 12 shows that errors are sizeable in our application. All error variances are significantly different from zero. Moreover, errors are larger for the test scores at age 7. For instance, in the one-factor specification the error variance represents 66%, 33% and 45% of the variances of the math test scores at age 7, 11 and 16, respectively. Overall, the ratio of the sum of squares of factor loadings to total variance is 60%. This suggests that overlooking error variables in the model can have severe consequences on the results. To check that, we re-estimated the factor loadings using JADE. We found that the second and third factors were essentially driven by the math and reading test scores at age 7, respectively. This is likely to be because the large errors in the test score at age 7 are wrongly interpreted as extra factors.

We also see from Table 12 that errors are generally contemporaneously positively correlated. The only exception is for the test scores at age 16 in the specification allowing for three factors ($K = 3$, last rows of the table).

Turning to factor loadings estimates, one sees that the one-factor specification weights all test scores similarly. The factor loadings estimates are very close to the ones we obtained using second-order moments only, by applying ordinary Factor Analysis. Moreover, they are very precisely estimated. A natural interpretation of this factor could be the child's general ability. From Table 13 we see that it is positively correlated with years of education (.50) and log wages (.30), and that it is equally distributed among boys and girls.

Allowing for a second factor yields a rather different picture, as none of the two factors is similar to the one estimated in the one-factor specification. The first factor is correlated with scores in reading and mathematics, the correlation being stronger with math. This factor is positively skewed, and presents negative excess kurtosis. In contrast, the second factor is correlated to reading test scores, but has small or zero correlation with the scores in mathematics. Contrary to the first one, this factor is both negatively skewed and leptokurtic. Moreover, the first and second factors account for 45% and 19% of the total variance, while errors account for 36%. Separating these two components requires to use third and fourth-order moments of the data, in order to fix the rotation matrix. This explains why standard errors are rather large compared to the one-factor specification. However, we remark that the estimates are still precise.

These results are consistent with the existence of different components of ability. Columns 2 and 3 in Table 13 show that the first factor is strongly related to math test scores, while the second only determines reading and verbal ability. Moreover, the first factor is strongly correlated with education and the log hourly wage (.45 and .30), while the second is less strongly correlated with education (.08) and is uncorrelated with the

Table 12: Model estimates

	Factor loadings			Error covariances						
ONE FACTOR										
Math (7)	14.3 (.26)	-	-	390 (5.7)	38.2 (3.8)	0	0	0	0	0
Reading (7)	15.5 (.22)	-	-	38.2 (3.8)	227 (4.2)	0	0	0	0	0
Math (11)	21.5 (.16)	-	-	0	0	219 (3.9)	43.3 (2.7)	69.7 (3.6)	0	0
Reading (11)	14.1 (.15)	-	-	0	0	43.3 (2.7)	114 (2.5)	38.1 (2.3)	0	0
Verbal (11)	18.5 (.17)	-	-	0	0	69.7 (3.6)	38.1 (2.3)	171 (3.5)	0	0
Math (16)	17.3 (.18)	-	-	0	0	0	0	0	228 (3.3)	8.58 (2.3)
Reading (16)	15.2 (.18)	-	-	0	0	0	0	0	8.58 (2.3)	101 (2.2)
Skewness	−.239 (.020)	-	-	-	-	-	-	-	-	-
Ex. kurtosis	−.888 (.029)	-	-	-	-	-	-	-	-	-
TWO FACTORS										
Math (7)	13.6 (.35)	4.61 (.48)	-	378 (6.0)	59.3 (4.0)	0	0	0	0	0
Reading (7)	10.8 (.44)	11.3 (.55)	-	59.3 (4.0)	230 (7.3)	0	0	0	0	0
Math (11)	22.3 (.35)	6.71 (.56)	-	0	0	118 (6.1)	17.3 (2.4)	45.6 (3.0)	0	0
Reading (11)	11.0 (.32)	9.92 (.32)	-	0	0	17.3 (2.4)	90.1 (2.3)	15.3 (2.1)	0	0
Verbal (11)	14.6 (.43)	12.1 (.46)	-	0	0	45.6 (3.0)	15.3 (2.1)	156 (3.4)	0	0
Math (16)	18.4 (.34)	4.24 (.50)	-	0	0	0	0	0	137 (9.3)	3.22 (3.5)
Reading (16)	11.1 (.35)	11.0 (.40)	-	0	0	0	0	0	3.42 (3.5)	85.5 (5.6)
Skewness	.600 (.047)	−1.71 (.087)	-	-	-	-	-	-	-	-
Ex. kurtosis	−1.29 (.045)	1.24 (.28)	-	-	-	-	-	-	-	-
THREE FACTORS										
Math (7)	13.6 (.25)	5.43 (.60)	−4.03 (.92)	363 (7.0)	16.4 (4.4)	0	0	0	0	0
Reading (7)	10.9 (.26)	13.4 (.90)	−6.64 (1.5)	16.4 (4.4)	141 (8.0)	0	0	0	0	0
Math (11)	22.4 (.21)	5.30 (.45)	−1.83 (1.1)	0	0	128 (5.1)	37.1 (2.9)	60.1 (3.1)	0	0
Reading (11)	11.0 (.25)	8.77 (.27)	2.06 (1.2)	0	0	37.1 (2.9)	104 (1.9)	37.2 (2.2)	0	0
Verbal (11)	14.8 (.28)	10.7 (.39)	−1.47 (1.4)	0	0	60.1 (3.1)	37.2 (2.2)	175 (3.7)	0	0
Math (16)	18.8 (.28)	4.07 (.41)	3.79 (.94)	0	0	0	0	0	114 (3.6)	−41.6 (2.1)
Reading (16)	12.2 (.37)	10.9 (.65)	7.05 (1.4)	0	0	0	0	0	−41.6 (2.1)	15.2 (1.7)
Skewness	.552 (.036)	−1.65 (.069)	.009 (.91)	-	-	-	-	-	-	-
Ex. kurtosis	−1.28 (.040)	1.28 (.21)	.520 (1.04)	-	-	-	-	-	-	-

Table 13: Correlation of the predicted factors with several variables

	One factor	Two factors		Three factors		
Math (7)	.682 (.008)	.685 (.031)	.035 (.048)	.662 (.018)	.070 (.059)	−.504 (.030)
Reading (7)	.751 (.005)	.352 (.034)	.654 (.033)	.367 (.016)	.685 (.054)	−.456 (.072)
Math (11)	.914 (.002)	.861 (.014)	.145 (.023)	.899 (.0072)	.088 (.020)	−.116 (.030)
Reading (11)	.842 (.004)	.538 (.022)	.521 (.022)	.593 (.017)	.476 (.019)	.135 (.063)
Verbal (11)	.877 (.003)	.575 (.025)	.521 (.022)	.640 (.015)	.452 (.020)	−.104 (.057)
Math (16)	.821 (.004)	.867 (.016)	−.011 (.024)	.870 (.013)	−.015 (.018)	.144 (.026)
Reading (16)	.821 (.004)	.492 (.023)	.555 (.024)	.356 (.021)	.542 (.030)	.291 (.067)
Years educ.	.494 (.010)	.454 (.013)	.078 (.012)	.470 (.011)	.065 (.011)	.110 (.012)
Log monthly wage	.261 (.015)	.293 (.015)	−.048 (.015)	.292 (.017)	−.046 (.017)	.093 (.013)
Log hourly wage	.281 (.017)	.290 (.015)	−.011 (.013)	.293 (.020)	−.014 (.016)	.081 (.014)
Female dummy	−.002 (.011)	−.136 (.012)	.203 (.011)	−.119 (.014)	.190 (.011)	−.128 (.011)

Note: bootstrapped standard errors in parentheses.

log hourly wage. This suggests that the second component of ability does not increase labor productivity. Lastly, girls are more likely to be endowed with the second factor, the negative correlation with the log monthly wage indicating that it is negatively associated with labor market participation.

Notice that, given $L = 7$ and $J = 16$, the bound on the number of factors that can be identified if only second-order moments are used in the prewhitening step of the algorithm is:

$$K = \frac{2L + 1 - \sqrt{(2L+1)^2 - 8J}}{2} \approx 2.58.$$

Hence, in order to identify a third factor, higher-order data moments are required in the first step. Adding a third factor, we remark that the first two factors remain unchanged: both factor loadings and moments are very similar to their values in the two-factor specification. This result confirms that the first two factors represent true dimensions of ability. As shown by Tables 12 and 13, the third factor puts positive weights on later test

scores (age 16) and negative weights on earlier ones (age 7). Moreover, it accounts for an additional 4% of total variance. This factor shows some excess kurtosis, though badly estimated, and it is positively correlated with education and log wages, albeit less so than the first factor (the correlation is .10 with education and the two log wage measures). Lastly, being a girl is negatively associated with this factor. We interpret the third factor as reflecting heterogeneous learning slopes. It allows to distinguish children who learn more at the beginning (age 7) or at the end (age 16) of their schooling career.

To conclude, this application shows that our algorithm succeeds in identifying three interpretable test score factors. A first dimension of children's ability reflects mathematical skills. It has a high positive return in terms of education and wages. The second dimension of ability is only correlated to reading and verbal test scores. It contributes a little to education, but does not increase labor market productivity. Moreover, it is more frequent among girls. The third dimension reflects the learning slopes of children. This last factor accounts for a small part of total variance, and has positive returns on education and wages.

7 Conclusion

The recent literature on Independent Component Analysis (ICA) has produced several methods able to deal with noise-free, linear independent factor models with up to $K = L$ factors. In this paper we have developed an algorithm that robustifies one of the most popular ICA algorithms, Cardoso and Souloumiac's (1993) JADE, when measurement error cannot be neglected. We have constructed a two-stage consistent estimator for noisy ICA with clustered errors, quasi-JADE. In the prewhitening step, error moments are estimated from second to fourth-order moments of the data, while in the source separation step JADE is applied to the whitened cumulant matrices.

Monte Carlo results are encouraging. For sufficiently non symmetric and/or kurtotic

data, we obtain small biases and precise estimates, even in relatively small samples. Moreover, the application to test scores shows that allowing for noise can be very important in practical situations. This suggests that quasi-JADE can be a valid alternative to existing methods in traditional applications of ICA, like signal processing, where it can be used in place of noise-free methods. Moreover, in situations where factor analysis is widely used (macroeconomics, finance, psychometrics) quasi-JADE provides a consistent way to fix the rotation matrix.

In the future, we plan to pursue two directions of research. First, we have shown that quasi-JADE can deal with a class of overcomplete ICA models. More work is needed for the general overcomplete case. The second direction of research concerns the extension of the method of this paper to the case of a very large number of measurements. Bai and Ng (2002) and Bai (2003) provide extensive analyses of the PCA estimator in this case. Financial and macroeconomic applications motivate the need to extend ICA methods in this direction.

Finally, once factor loadings have been estimated, it remains to estimate the distribution of factors and errors. This is done in a companion paper (see Bonhomme and Robin, 2006).

APPENDIX

A Mathematical proofs

We start with some notation. For a n-by-m matrix A, we denote as $A[I,J]$ the submatrix of rows $i \in I$ and columns $j \in J$, for $I \subseteq \{1,...,n\}$ and $J \subseteq \{1,...,m\}$. If $I = \{1,...,n\}$ or $J = \{1,...,m\}$, we write $A[\cdot, J]$ and $A[I, \cdot]$.

So, in particular, matrix $Q_{\mathcal{J}}$ can be equivalently written as $Q[\mathcal{J}, \cdot]$, and $\Lambda_{\mathcal{I}_\ell}$ as $\Lambda[\mathcal{I}_\ell, .]$.

A.1 Proof of Lemma 1

Let $(\ell, m) \in \mathcal{J}$. As $U_\ell \perp\!\!\!\perp U_m$, we have $\text{Cov}(U_\ell, U_m) = 0$. In addition: $U_\ell = \Pi_\ell^{\mathrm{T}} \varepsilon \perp\!\!\!\perp U_m = \Pi_m^{\mathrm{T}} \varepsilon$, where Π_ℓ^{T} is the ℓth row of matrix Π. It follows from Darmois' theorem (e.g., Comon, 1994, p.306) that for all $h \in \{1,...,H\}$ either ε_h is Gaussian or $\pi_{\ell h}\pi_{mh} = 0$. In either case:

$$\pi_{\ell h}\pi_{mh}\kappa_3(\varepsilon_h) = \pi_{\ell h}\pi_{mh}\kappa_4(\varepsilon_h) = 0.$$

The conclusion comes from the cumulant identities:

$$\begin{aligned} \text{Cum}(U_i, U_\ell, U_m) &= \sum_{h=1}^{H} \pi_{ih}\pi_{\ell h}\pi_{mh}\kappa_3(\varepsilon_h), \\ \text{Cum}(U_i, U_j, U_\ell, U_m) &= \sum_{h=1}^{H} \pi_{ih}\pi_{jh}\pi_{\ell h}\pi_{mh}\kappa_4(\varepsilon_h). \end{aligned}$$

A.2 Proof of Theorem 2

To simplify the exposition, let us define $\sigma_Y \equiv \text{vech}(\Sigma_Y)$, $\gamma_Y(\ell) \equiv \text{vech}(\Gamma_Y(\ell))$, and $\omega_Y(\ell, m) \equiv \text{vech}(\Omega_Y(\ell, m))$, with similar notation for σ_U, $\gamma_U(\ell)$ and $\omega_U(\ell, m)$. Let also

$$\mathcal{J}^c = \left\{ (\ell, m) \in \{1,...,L\}^2, \ell \leq m \right\} \setminus \mathcal{J}.$$

Remark that σ_U, $\gamma_U(\ell)$ and $\omega_U(\ell, m)$ have zero entries in positions $(i,j) \in \mathcal{J}$. Construct vectors $\sigma_U[\mathcal{J}^c]$, $\gamma_U(\ell)[\mathcal{J}^c]$ and $\omega_U(\ell, m)[\mathcal{J}^c]$ by dropping the zero entries. Let also $B[\mathcal{J}^c, \cdot]$ be the submatrix obtained by selecting the rows of B indexed by couples $(\ell, m) \notin \mathcal{J}$. Equations (8), (9) and (10) imply

$$B^{\mathrm{T}}\sigma_Y = B[\mathcal{J}^c, \cdot]^{\mathrm{T}} \sigma_U[\mathcal{J}^c], \tag{A1}$$

$$B^{\mathrm{T}}\gamma_Y(\ell) = B[\mathcal{J}^c, \cdot]^{\mathrm{T}} \gamma_U(\ell)[\mathcal{J}^c], \quad \forall \ell, \tag{A2}$$

$$B^{\mathrm{T}}\gamma_Y(\ell, m) = B[\mathcal{J}^c, \cdot]^{\mathrm{T}} \omega_U(\ell, m)[\mathcal{J}^c], \quad \forall (\ell, m). \tag{A3}$$

We shall show that matrix $B[\mathcal{J}^c, \cdot]$ has full row rank, which will prove the identification of error moments. To proceed, remark that $B[\mathcal{J}^c, \cdot]$ has $\frac{L(L+1)}{2} - J$ rows and $\frac{L(L+1)}{2} - K$ columns. If $J \geq K$, $B[\mathcal{J}^c, \cdot]$ has more columns than rows. Let $r = \text{rank}\,(B[\mathcal{J}^c, \cdot])$.

Suppose that $r < \frac{L(L+1)}{2} - J$. There exists a $\left(\frac{L(L+1)}{2} - K\right)$-by-$\left(\frac{L(L+1)}{2} - K - r\right)$ matrix A, full column rank, such that $B[\mathcal{J}^c, \cdot]\, A = 0$. As both B and A have full column rank, BA has full column rank, hence $B[\mathcal{J}, \cdot]\, A$ necessarily has full column rank $\frac{L(L+1)}{2} - K - r$, with

$$\frac{L(L+1)}{2} - K - r > J - K. \tag{A4}$$

Moreover, as $Q^{\mathrm{T}} B = 0$ by construction,

$$0 = Q^{\mathrm{T}} BA = Q[\mathcal{J}, \cdot]^{\mathrm{T}}\, B[\mathcal{J}, \cdot]\, A,$$

Now, $Q[\mathcal{J}, \cdot]$ has J rows and K columns. It has full column rank, so its null space has dimension $J - K$. This contradicts condition (A4) on the rank of $B[\mathcal{J}, \cdot]\, A$. Hence, $r = \frac{L(L+1)}{2} - J$ and matrix $B[\mathcal{J}^c, \cdot]$ therefore must have full row rank.

This ends the proof of Theorem 2.

A.3 Proof of Theorem 3

Let us define $\mathcal{I}_\ell^c = \{m \in \{1, ..., L\}, m \notin \mathcal{I}_\ell\}$, for all $\ell \in \{1, ..., L\}$, that is,

$$\mathcal{I}_\ell^c = \{m \in \{1, ..., L\}, \ell \leq m \text{ and } (\ell, m) \in \mathcal{J}^c\}.$$

1. We first show that, for all ℓ, $C[\mathcal{I}_\ell^c, \cdot]$ has full row rank in the same way as in the proof of Theorem 2.

Matrix $C[\mathcal{I}_\ell^c, \cdot]$ has $L - I_\ell$ rows and $L - K$ columns. As, by assumption, $\Lambda[\mathcal{I}_\ell, \cdot]$ has rank K and dimensions I_ℓ-by-K, $I_\ell \geq K$. Suppose that $r = \text{rank}\,(C[\mathcal{I}_\ell^c, \cdot]) < L - I_\ell$. There exists a full column rank, $(L-K)$-by-$(L-K-r)$ matrix A, such that $C[\mathcal{I}_\ell^c, \cdot]\, A = 0$. Both A and C having full column rank, CA has also full column rank. Hence, $C[\mathcal{I}_\ell, \cdot]\, A$ has full colum rank $L - K - r$.

Moreover, $C^{\mathrm{T}}\Lambda = 0$. Hence,

$$0 = \Lambda^{\mathrm{T}} CA = \Lambda[\mathcal{I}_\ell, \cdot]^{\mathrm{T}}\, C[\mathcal{I}_\ell, \cdot]\, A.$$

By assumption, $\Lambda[\mathcal{I}_\ell, \cdot]$ is full column rank K. Its null space thus has dimension $I_\ell - K$. Therefore $C[\mathcal{I}_\ell, \cdot]\, A$ cannot have a rank greater than $I_\ell - K$:

$$L - K - r \leq I_\ell - K.$$

Hence $r \geq L - I_\ell$, which contradicts the assumption.

2. Now, applying the vech operator to (11), (12), (13) shows that error cumulants satisfy the linear system (A1), (A2), (A3) with, in place of $B\left[\mathcal{J}^c, \cdot\right]$, the block diagonal matrix

$$D \equiv \operatorname{diag}\left(C\left[\mathcal{I}_1^c, \cdot\right], ..., C\left[\mathcal{I}_L^c, \cdot\right]\right).$$

As $C\left[\mathcal{I}_\ell^c, \cdot\right]$ has full row rank for all ℓ, it follows that D has also full row rank.

This ends the proof of Theorem 3.

A.4 Proof of Theorem 4

To prove Theorem 4, we first prove the following lemma giving conditions under which the joint eigenvectors of a set of matrices is uniquely defined (up to sign and permutation).

Lemma 2 *Let K and L be any integers. Let $A_1, ..., A_L$ be K-by-K matrices. Suppose that there exist $x^k = \left(x_1^k, ..., x_L^k\right)^{\mathrm{T}} \in \mathbb{R}^L$ and $v^k \in \mathbb{R}^K$, $v_k \neq 0$, $k = 1, ..., K+1$, solutions to the joint diagonalization problem:*

$$x_\ell^k v^k = A_\ell v^k, \quad \forall \ell = 1, ..., L.$$

Assume that the set $\left\{v^1, ..., v^K\right\}$ is linearly independent, that all v_k, $k = 1, ..., K+1$, have norm one, and that $x^k \neq x^{k'}$ for all $(k, k') \in \{1, ..., K\}^2$, $k \neq k'$. Then there exists $k \in \{1, ..., K\}$ such that $v^{K+1} = \pm v^k$.

Proof. Since $\left\{v^1, ..., v^K\right\}$ is a basis of $\mathbb{R}^K$, there exists $c = (c_1, ..., c_K) \neq 0$ such that $v^{K+1} = c_1 v^1 + ... + c_K v^K$. Then, for all $\ell = 1, ..., L$,

$$\sum_{k=1}^K c_k x_\ell^k v^k = \sum_{\ell=1}^K c_k A_\ell v^k = A_\ell \sum_{\ell=1}^K c_k v^k = A_\ell v^{K+1} = x_\ell^{K+1} v^{K+1} = x_\ell^{K+1} \left(\sum_{k=1}^K c_k v^k\right).$$

As $(v^1, ..., v^K)$ is linearly independent, it follows from the last equality that:

$$c_k x_\ell^k = c_k x_\ell^{K+1},$$

for all (k, ℓ). Hence, for all k:

$$c_k x^k = c_k x^{K+1}.$$

As $c \neq 0$, there exists k such that $c_k \neq 0$. For this k: $x^k = x^{K+1}$. Moreover, as $x^k \neq x^{k'}$ for all $k' \neq k$ in $\{1, ..., K\}$, it follows that $c_{k'} = 0$ for all $k' \neq k$. Hence

$$v^{K+1} = c_k v^k.$$

As both v^k and v^{K+1} have norm one, $c_k = \pm 1$. The result follows. ■

The proof of Theorem 4 easily follows.

Fourth-order moments. Second and fourth-order cumulant restrictions (3)-(5) yield:

$$\widetilde{\Omega}_Y(\ell, m) = \Lambda D_4 \operatorname{diag}(\Lambda_\ell \odot \Lambda_m) \Lambda^{\mathrm{T}}, \quad (\ell, m) \in \overline{\Delta}_L, \tag{A5}$$

$$\widetilde{\Sigma}_Y = \Lambda\Lambda^{\mathrm{T}}. \tag{A6}$$

Let $\widetilde{\Lambda}$ be another value satisfying restrictions (A5)-(A6). We show that under the conditions of Theorem 4, there necessarily exists a sign-permutation matrix S such that $\widetilde{\Lambda} = \Lambda S$.

Λ having full column rank K, and $\widetilde{\Sigma}_Y$ being positive definite, there exists a unique orthonormal L-by-K matrix W ($W^{\mathrm{T}}W = I_K$) and a unique K-by-K diagonal, positive matrix D such that $\widetilde{\Sigma}_Y = WDW^{\mathrm{T}}$. Let $P = D^{-1/2}W^{\mathrm{T}}$. Then $V = P\Lambda$ is a matrix of joint orthonormal eigenvectors ($VV^{\mathrm{T}} = I_K$) of

$$P\widetilde{\Omega}_Y(\ell, m) P^{\mathrm{T}} = P\Lambda D_4 \operatorname{diag}(\Lambda_\ell \odot \Lambda_m) \Lambda^{\mathrm{T}} P^{\mathrm{T}}, \quad \ell \leq m.$$

In general, there can be infinitely many joint eigenvectors to a set of matrices if all matrices have multiple roots. However, Lemma 2 shows that the problem of diagonalizing matrices $P\widetilde{\Omega}_Y(\ell, m) P^{\mathrm{T}}$ has a unique solution up to column sign and permutation if for all $(k, k') \in \{1, ..., K\}^2$, $k \neq k'$, there exists $\ell \leq m$ such that

$$\lambda_{\ell k}\lambda_{mk}\kappa_4(X_k) \neq \lambda_{\ell k'}\lambda_{mk'}\kappa_4(X_{k'}).$$

As either $\kappa_4(X_k) \neq 0$ or $\kappa_4(X_{k'}) \neq 0$, and as any two columns of Λ are linearly independent, this condition is always satisfied. It follows that V is uniquely defined, up to column sign and permutation.

Now, the true Λ necessarily verifies:

$$\Lambda = \Lambda(P\Lambda)^{\mathrm{T}}(P\Lambda) = \Lambda\Lambda^{\mathrm{T}}P^{\mathrm{T}}P\Lambda = \widetilde{\Sigma}_Y P^{\mathrm{T}}P\Lambda = WP\Lambda = WV.$$

It is thus unique as V is unique.

Third-order moments. The same argument applies to third-order cumulant matrices $\widetilde{\Gamma}_Y(\ell)$. Indeed, in the noise-free case third-order restrictions (4) become

$$\widetilde{\Gamma}_Y(\ell) = \Lambda D_3 \operatorname{diag}(\Lambda_\ell) \Lambda^{\mathrm{T}}, \quad \ell \in \{1, ..., L\}.$$

In this case, Lemma 2 shows that the common eigenvectors corresponding to eigenvalues $D_3 \operatorname{diag}(\Lambda_\ell)$ are uniquely determined up to column sign and permutation if for all $(k, k') \in \{1, ..., K\}^2$, $k \neq k'$, there exists $\ell \in \{1, ..., L\}$ such that

$$\lambda_{\ell k} \kappa_3 (X_k) \neq \lambda_{\ell k'} \kappa_3 (X_{k'}) .$$

As before, this condition is always satisfied.

Third and fourth-order moments. The proof is almost identical to the two previous ones. With $\widetilde{\Omega}_Y (\ell, m)$ and $\widetilde{\Gamma}_Y (\ell)$ together, eigenvectors are identified if for all $(k, k') \in \{1, ..., K\}^2$, $k \neq k'$, there exists (ℓ, m) such that

$$\lambda_{\ell k} \lambda_{mk} \kappa_4 (X_k) \neq \lambda_{\ell k'} \lambda_{mk'} \kappa_4 (X_{k'}) ,$$

or there exists $\ell \in \{1, ..., L\}$ such that

$$\lambda_{\ell k} \kappa_3 (X_k) \neq \lambda_{\ell k'} \kappa_3 (X_{k'}) .$$

As one of the four moments $\kappa_3 (X_k)$, $\kappa_3 (X_{k'})$, $\kappa_4 (X_k)$ and $\kappa_4 (X_{k'})$ is non zero, it follows from the assumptions on Λ that this condition is always satisfied.

B The JADE algorithm

Let $\mathcal{A} = \{A_k, k = 1...K\}$ a set of real, symmetric, L-by-L matrices. Let us define the function:

$$\operatorname{off}(A) = \sum_{i \neq j} a_{ij}^2,$$

for all $A = [a_{ij}]$. Then joint diagonalization of $\mathcal{A}$ is achieved by minimizing

$$\sum_{k=1}^{K} \operatorname{off}(U A_k U^{\mathrm{T}}), \tag{B7}$$

with respect to U orthogonal.

Let $\theta \in [-\pi, \pi]$, let $(i, j) \in \{1, ..., L\}^2$ and let $R_{ij}(\theta)$ be the L-by-L matrix equal to zero everywhere except at the (i, i), (i, j), (j, i) and (j, j) entries where it is equal to:

$$\begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix}.$$

Let $i \neq j$, and let us define:

$$O_{i,j}(\theta) = \sum_{k=1}^{K} \operatorname{off}\left(R_{ij}(\theta) A_k R_{ij}(\theta)^{\mathrm{T}}\right).$$

Lastly, let $h_{i,j}(A) = (a_{ii} - aij, a_{ij} + a_{ji})$, and let:

$$G_{i,j} = \sum_{k=1}^{K} h_{i,j}^{\mathrm{T}}(A_k) h_{i,j}(A_k) = (g_{ij})_{i,j=1,2}\,.$$

Cardoso and Souloumiac (1996) show that θ_0 such that:

$$\cos(\theta_0) = \sqrt{\frac{x+r}{2r}}, \quad \sin(\theta_0) = \sqrt{\frac{y}{2r(x+r)}},$$

where $x = g_{11} - g_{22}$, $y = g_{12} + g_{21}$ and $r = \sqrt{x^2 + y^2}$, minimizes $O_{i,j}(\theta)$.

This closed-form expression for θ_0 allows to minimize (B7) by the following algorithm:

1. Start with $U(0) = I_L$.
2. Begin loop on step s.
3. Begin loop on (i, j).
4. Compute $G_{i,j}$.
5. Compute θ_0.
6. If θ_0 is different enough from zero, continue. Else stop.
7. Compute $R_{ij}(\theta_0) A_k R_{ij}(\theta_0)^{\mathrm{T}}$ and modify $\mathcal{A}$ consequently.
8. Update $U(s)$ as $U(s+1) = R_{ij}(\theta_0) U(s)$.
9. End loop on (i, j).
10. End loop on s.

C Asymptotic theory of the JADE estimator

First-order conditions. The JADE estimator solves

$$\widehat{V} = \arg \min_{V \in \mathcal{O}_K} \sum_{s=1}^{S} \mathrm{off}(V^{\mathrm{T}} \widehat{A}_s V).$$

The Lagrangian associated with the minimization problem is:

$$\begin{aligned} \mathcal{L}(V, \gamma) &= \sum_{s=1}^{S} \mathrm{off}(V^{\mathrm{T}} \widehat{A}_s V) + \gamma^{\mathrm{T}} \,\mathrm{vec}\left(V^{\mathrm{T}} V - I_K\right), \\ &= \sum_{s} \sum_{m \neq k} (v_k^{\mathrm{T}} \widehat{A}_s v_m)^2 + \sum_{k} \gamma_{kk} (v_k^{\mathrm{T}} v_k - 1) + \sum_{m \neq k} \gamma_{mk} v_k^{\mathrm{T}} v_m, \end{aligned}$$

where γ is a vector of K^2 Lagrange multipliers γ_{mk}, and v_k is the kth column of matrix V.

Differentiating the Lagrangian with respect to v_ℓ, for $\ell = 1...K$, yields:

$$\frac{\partial \mathcal{L}\left(\widehat{V}, \widehat{\gamma}\right)}{\partial v_\ell} = 2\sum_s \sum_{k\neq\ell} (\widehat{v}_k^{\mathrm{T}} \widehat{A}_s \widehat{v}_\ell)\widehat{A}_s \widehat{v}_k + 2\widehat{\gamma}_{\ell\ell}\widehat{v}_\ell + \sum_{k\neq\ell} \widehat{\gamma}_{k\ell}\widehat{v}_k = 0.$$

Then, multiplying this equation by $\widehat{v}_m^{\mathrm{T}}$, for $m \neq \ell$, gives:

$$2\sum_s \sum_{k\neq\ell} (\widehat{v}_k^{\mathrm{T}} \widehat{A}_s \widehat{v}_\ell)\widehat{v}_m^{\mathrm{T}}\widehat{A}_s \widehat{v}_k + \widehat{\gamma}_{m\ell} = 0.$$

Using that $\widehat{\gamma}_{m\ell} = \widehat{\gamma}_{\ell m}$ by symmetry, it follows that

$$\sum_s \sum_{k\neq\ell} (\widehat{v}_k^{\mathrm{T}} \widehat{A}_s \widehat{v}_\ell)\widehat{v}_m^{\mathrm{T}}\widehat{A}_s \widehat{v}_k = \sum_s \sum_{k\neq m} (\widehat{v}_k^{\mathrm{T}} \widehat{A}_s \widehat{v}_m)\widehat{v}_\ell^{\mathrm{T}}\widehat{A}_s \widehat{v}_k,$$

or, equivalently, as $\widehat{A}_s$ is symmetric for all s:

$$\sum_s \widehat{v}_\ell^{\mathrm{T}}\widehat{A}_s \left(\sum_{k\neq\ell} \widehat{v}_k\widehat{v}_k^{\mathrm{T}}\right) \widehat{A}_s\widehat{v}_m = \sum_s \widehat{v}_m^{\mathrm{T}}\widehat{A}_s \left(\sum_{k\neq m} \widehat{v}_k\widehat{v}_k^{\mathrm{T}}\right) \widehat{A}_s\widehat{v}_\ell.$$

Then, as $\sum_{k=1}^K \widehat{v}_k\widehat{v}_k^{\mathrm{T}} = \widehat{V}\widehat{V}^{\mathrm{T}} = I_K$ we obtain

$$\sum_s \widehat{v}_\ell^{\mathrm{T}}\widehat{A}_s \left(I_K - \widehat{v}_\ell\widehat{v}_\ell^{\mathrm{T}}\right) \widehat{A}_s\widehat{v}_m = \sum_s \widehat{v}_m^{\mathrm{T}}\widehat{A}_s \left(I_K - \widehat{v}_m\widehat{v}_m^{\mathrm{T}}\right) \widehat{A}_s\widehat{v}_\ell,$$

which we write after rearranging:

$$\sum_s \widehat{v}_\ell^{\mathrm{T}}\widehat{A}_s\widehat{v}_m \left(\widehat{v}_m^{\mathrm{T}}\widehat{A}_s\widehat{v}_m - \widehat{v}_\ell^{\mathrm{T}}\widehat{A}_s\widehat{v}_\ell\right) = 0. \tag{C8}$$

Equation (C8) holds for all $\ell < m$. The JADE estimator $\widehat{V}$ solves these $K(K-1)/2$ non redundant equations, together with the $K(K+1)/2$ orthogonality constraints:

$$\widehat{v}_\ell^{\mathrm{T}}\widehat{v}_m = \delta_{\ell m}, \text{ for all } \ell \leq m.$$

Identification and consistency. Let $\widetilde{V} = (\widetilde{v}_1, ..., \widetilde{v}_K) \in \mathcal{O}_K$ be such that

$$\widetilde{V} = \arg\min_{V\in\mathcal{O}_K} \sum_{s=1}^S \mathrm{off}(V^{\mathrm{T}} A_s V).$$

Then, as: $\min_{V\in\mathcal{O}_K} \sum_{s=1}^S \mathrm{off}(V^{\mathrm{T}}A_sV) = 0$ at the true value, it follows that $\widetilde{V}^{\mathrm{T}}A_s\widetilde{V} = \widetilde{D}_s$ is diagonal for all s. As for all $k \neq m$ there exists $s \in \{1...S\}$ such that $d_{sk} \neq d_{sm}$, one can apply Theorem 2 to show that $\widetilde{V}$ is equal to the true V, up to column sign and permutation. This shows the identification of V. Consistency follows from classical arguments, as the parameter space $\mathcal{O}_K$ is compact.

Asymptotic distribution. A first-order Taylor expansion of (C8) around the true value V yields:

$$\sum_s^S v_m^{\mathrm{T}}\widehat{A}_s v_k\left(v_k^{\mathrm{T}}\widehat{A}_s v_k - v_m^{\mathrm{T}}\widehat{A}_s v_m\right)+\sum_s^S\left(v_k^{\mathrm{T}}\widehat{A}_s v_k - v_m^{\mathrm{T}}\widehat{A}_s v_m\right)\left(v_m^{\mathrm{T}}\widehat{A}_s(\widehat{v}_k - v_k)+v_k^{\mathrm{T}}\widehat{A}_s(\widehat{v}_m - v_m)\right)$$
$$+\sum_s^S v_m^{\mathrm{T}}\widehat{A}_s v_k\left(v_k^{\mathrm{T}}\widehat{A}_s(\widehat{v}_k - v_k) - v_m^{\mathrm{T}}\widehat{A}_s(\widehat{v}_m - v_m)\right) = o_p\left(N^{-1/2}\right).$$

As $\underset{N\to\infty}{\operatorname{plim}}\widehat{A}_s = A_s$ for all s, and as $v_k^{\mathrm{T}}A_s v_m = 0$ for all $k \neq m$, this yields:

$$\sum_s^S (d_{sk} - d_{sm})\, v_m^{\mathrm{T}}\left(\widehat{A}_s - A_s\right)v_k+\sum_s^S (d_{sk} - d_{sm})\left(v_m^{\mathrm{T}}A_s(\widehat{v}_k - v_k)+v_k^{\mathrm{T}}A_s(\widehat{v}_m - v_m)\right) = o_p\left(N^{-1/2}\right),$$

where $d_{sk} = v_k^{\mathrm{T}}A_s v_k$ are the diagonal elements of $V^{\mathrm{T}}A_s V$.

At this stage, it is convenient to define $\widehat{x}_{mk} \equiv v_m^{\mathrm{T}}(\widehat{v}_k - v_k)$. As $v_m^{\mathrm{T}}A_s = d_{sm}v_m^{\mathrm{T}}$, one has:

$$\sum_s^S (d_{sk} - d_{sm})\, v_m^{\mathrm{T}}\left(\widehat{A}_s - A_s\right)v_k + \sum_s^S (d_{sk} - d_{sm})\,(d_{sm}\widehat{x}_{mk} + d_{sk}\widehat{x}_{km}) = o_p\left(N^{-1/2}\right).$$

Now, a Taylor expansion of the orthogonality constraints yields:

$$\widehat{x}_{mk} + \widehat{x}_{km} = v_m^{\mathrm{T}}(\widehat{v}_k - v_k) + v_k^{\mathrm{T}}(\widehat{v}_m - v_m) = 0, \text{ for all } m, k.$$

Thus we have:

$$\sum_s^S (d_{sk} - d_{sm})^2\,\widehat{x}_{mk} = -\sum_s^S (d_{sk} - d_{sm})\, v_m^{\mathrm{T}}\left(\widehat{A}_s - A_s\right)v_k + o_p\left(N^{-1/2}\right). \tag{C9}$$

Let $\widehat{X} = V^{\mathrm{T}}\left(\widehat{V} - V\right)$. Then equation (C9) is equivalently written, in matrix form, as:

$$\operatorname{vec}\left(\widehat{X}\right) = -W\left(I_S \otimes V^{\mathrm{T}} \otimes V^{\mathrm{T}}\right)\left(\operatorname{vec}\left(\widehat{A}\right) - \operatorname{vec}(A)\right) + o_p\left(N^{-1/2}\right),$$

where W, A and $\widehat{A}$ have been defined in the text. Note that W is provided that $\sum_s^S (d_{sk} - d_{sm})^2 \neq 0$ for all $k \neq m$.

Then, as:

$$\operatorname{vec}\left(\widehat{X}\right) = \left(I_K \otimes V^{\mathrm{T}}\right)\left(\operatorname{vec}\left(\widehat{V}\right) - \operatorname{vec}(V)\right),$$

it follows that

$$N^{\frac{1}{2}}\left(\operatorname{vec}\left(\widehat{V}\right) - \operatorname{vec}(V)\right) = -\left(I_K \otimes V\right)W\left(I_S \otimes V^{\mathrm{T}} \otimes V^{\mathrm{T}}\right)N^{\frac{1}{2}}\left(\operatorname{vec}\left(\widehat{A}\right) - \operatorname{vec}(A)\right) + o_p(1),$$

from which

$$N^{\frac{1}{2}}\left(\operatorname{vec}(\widehat{V}) - \operatorname{vec}(V)\right) \underset{d}{\to} \mathcal{N}\left(0, \mathbb{V}_V\right),$$

where the expression of $\mathbb{V}_V$ is given by (20).

D Robin and Smith's (2000) rank test

Let $\widehat{B}$ be a root-N consistent estimator of a (p,q), $p \geq q$, matrix B, such that

$$N^{1/2}\,\mathrm{vec}\left(\widehat{B}-B\right) \xrightarrow{d} \mathcal{N}\left(0, \Sigma_{\mathrm{vec}\left(\widehat{B}\right)}\right),$$

where $\Sigma_{\mathrm{vec}\left(\widehat{B}\right)}$ is definite and $\mathrm{rank}\left(\Sigma_{\mathrm{vec}\left(\widehat{B}\right)}\right) = s$, $0 < s \leq pq$. Note that $s < \dim(V)$ because of the symmetry properties of Γ_Y and Ω_Y. Let $\widehat{\Sigma}_{\mathrm{vec}\left(\widehat{B}\right)}$ be a consistent estimate of $\Sigma_{\mathrm{vec}\left(\widehat{B}\right)}$. Let $\widehat{B} = \widehat{C}\widehat{D}\widehat{E}^{\mathrm{T}}$ be the singular value decomposition of $\widehat{B}$, where $\widehat{C}$ and $\widehat{E}$ are (p,p) and (q,q) orthogonal matrices and $\widehat{D}$ is a (q,p) diagonal matrix. Let $\widehat{d}_1 \geq ... \geq \widehat{d}_K$ denote the diagonal entries of $\widehat{D}^2$ (the eigenvalues of $\widehat{B}^{\mathrm{T}}\widehat{B}$). For a given null hypothesis: $H_0^r : K = r$, the statistics

$$\mathcal{CRT}_r \equiv N \sum_{i=r+1}^{q} \widehat{d}_i$$

has the same limiting distribution as $\sum_{i=1}^{\mathrm{T}} d_i^r Z_i^2$, where $d_1^r \geq ... \geq d_t^r$, $t \leq \min\{s, (p-r)(q-r)\}$, are the non-zero ordered eigenvalues of the matrix

$$(\widehat{E}_{q-r} \otimes \widehat{C}_{p-r})^{\mathrm{T}} \widehat{\Sigma}_{\mathrm{vec}\left(\widehat{B}\right)} (\widehat{E}_{q-r} \otimes \widehat{C}_{p-r}),$$

where $\widehat{E}_{q-r}$ and $\widehat{C}_{p-r}$ are the last $q-r$ and $p-r$ columns of $\widehat{E}$ and $\widehat{C}$, respectively, and $\{Z_i\}_{i=1}^{\mathrm{T}}$ are independent standard normal variates.

To estimate K, we apply the following procedure. Start with $r = 0$. Test H_0^1 against $\widetilde{H}_0^1 : K > 0$. If H_0^1 is rejected, test H_0^2 against $\widetilde{H}_0^2 : K > 1$. And so on until one accepts H_0^r against $\widetilde{H}_0^r : K > r$. The test p-values can be approximated by drawing many independent values of the limiting statistics $\sum_{i=1}^{\mathrm{T}} d_i^r Z_i^2$. This procedure delivers a consistent estimate of K if the asymptotic sizes α_N^r used for the sequential tests are such that $\alpha_N^r = o(1)$ and $-N^{-1}\ln\alpha_N^r = o(1)$.

References

[1] ANDERSON, T. W. (1963): "Asymptotic Theory for Principal Component Analysis," *Ann. Math. Stat.*, 34, 122-148.

[2] ANDERSON, T. W., and H. RUBIN (1956): "Statistical Inference in Factor Analysis," in *Proceedings of the Third Symposium in Mathematical Statistics and Probability*, Vol. 5. University of California press.

[3] ANSELIN, L. (2003): "Spatial Externalities, Spatial Multipliers, and Spatial Econometrics," *International Regional Science Review,* 26(2), 153-166.

[4] ATTIAS, H. (1999), "Independent Factor Analysis," *Neural Computation*, ;11:803-851.

[5] BAI, J. (2003): "Inferential Theory for Factor Models of Large Dimensions," *Econometrica,* 71, 135-171.

[6] BAI, J., and S. NG (2002): "Determining the Number of Factors in Approximate Factor Models," *Econometrica,* 70, 191-221.

[7] BECKMANN, C.F., and S.M. SMITH (2004): "Probabilistic Independent Component Analysis for Functional Magnetic Resonance Imaging, " *IEEE Trans. on Medical Imaging*, 23(2), 137-152.

[8] BLANCHARD, O.J., and D. QUAH (1989): "The Dynamic Effects of Aggregate Demand and Supply Disturbances, " *American Economic Review*, 79(4), 655-673.

[9] BONHOMME, S., and J.M. ROBIN (2006): "Generalized Nonparametric Deconvolution with an Application to Earnings Dynamics," *mimeo.*

[10] CARDOSO, J.-F. (1999): "High-order contrasts for independent Component Analysis," *Neural Computation,* 11, 157-192.

[11] CARDOSO, J.-F., and A. SOULOUMIAC (1993): "Blind Beamforming for Non-Gaussian Signals," *IEE-Proceedings-F,* 140, 362-370.

[12] CARDOSO, J.-F., and A. SOULOUMIAC (1996): "Jacobi Angles for Simultaneous Diagonalization," *SIAM J. Mat. An. Appl.,* 17, 161-164.

[13] CARDOSO, J.-F., and D.-T. PHAM (2004): "Optimization Issues in Noisy Gaussian ICA," *Proc ICA 2004,* Granada, Spain.

[14] CARNEIRO, P., K. HANSEN, and J.J. HECKMAN (2003), "Estimating Distributions of Treatment Effects with an Application to the Returns to Schooling and Measurement of the Effects of Uncertainty on College Choice," *International Economic Review*, 44(2): 361-422.

[15] CHEN, A., and P.J. BICKEL (2005): "Consistent Independent Component Analysis and Prewhitening," *IEEE Trans. on Signal Processing,* 53(10), 3625-3632.

[16] COMON, P. (1994): "Independent Component Analysis, a New Concept?," *Signal Processing*, 36(3), 287-314.

[17] COMON, P. (2004): "Blind Identification and Source Separation in 2×3 Under-determined Mixtures," *IEEE Trans. Signal Processing*, 11-22.

[18] CRAGG, J. G. (1997): "Using Higher Moments to Estimate the Simple Errors-in-Variables Model," *RAND Journal of Economics*, 28, S71-S91.

[19] DAGENAIS, M. G., and D. L. DAGENAIS (1997): "Higher Moment Estimators for Linear Regression Models with Errors in Variables," *Journal of Econometrics*, 76, 193-221.

[20] DAVIES, M. (2004): "Identifiability Issues in Noisy ICA," *IEEE Signal Processing Letters*, 11(5), 470-473.

[21] ERICKSON, T., and T. WHITED (2002): "Two-Step GMM Estimation of the Error-in-Variables Model Using High-Order Moments," *Econometric Theory*, 18, 776-799.

[22] ERIKSSON, J., and V. KOIVUNEN (2003): "Identifiability and separability of linear ICA models revisited," *4th International Symposium on ICA and Blind Signal Separation*, 23-27.

[23] FLURY, B. (1986): "Asymptotic Theory for Common Principal Component Analysis," *Annals of Statistics*, 14, 418-430.

[24] FORNI, M., and L. REICHLIN (1998): "Let's Get Real: A Factor Analytical Approach to Disaggregated Business Cycle Dynamics," *Review of Economic Studies*, 65, 453-473.

[25] GEARY, R. C. (1942): "Inherent Relations Between Random Variables," *Proc. Royal Irish Academy*, 47, 63-76.

[26] HECKMAN, J.J., J. STIXRUD and S. URZUA (2006): "The Effect of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior," *Journal of Labor Economics*, 24(39), 411-482.

[27] HYVARINEN, A., J. KARHUNEN and E. OJA (2001): *Independent Component Analysis*, John Wiley & Sons, New York.

[28] HYVARINEN, A., and E. OJA (2001): "A Fast Fixed Point Algorithm for Independent Component Analysis," *Neural Computation*, 9(7), 1483-1492.

[29] IKEDA, S., and K. TOYAMA (2000): "Independent Component Analysis for Noisy Data–MEG Data Analysis," *Neural Networks*, 13(10), 1063-1074.

[30] JENNRICH, R.I., and N. TRENDAFILOV (2005): "Independent Component Analysis as a Rotation Method: A Very Different Solution to Thurstone's Box Problem," *British Journal of Mathematical and Statistical Psychology*, 58, 199-208.

[31] KAISER, H.F. (1958): "The Varimax Criterion for Analytic Rotation in Factor Analysis," *Psychometrika*, 23(3), 187-200.

[32] KANO, Y., MIYAMOTO, Y., and S. SHIMIZU (2003): "Factor Rotation and ICA," *4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA 2003)*, 101-105.

[33] LEWBEL, A. (1997): "Constructing Instruments for Regressions with Measurement Error When No Additional Data are Available, with an Application to Patents and R&D," *Econometrica*, 65, 1201-1213.

[34] MOOIJAART, A. (1985): "Factor Analysis for Non-Normal Variables," *Psychometrika*, 50(3), 323-342.

[35] MOULINES, E. J.-F. CARDOSO and E. GASSIAT (1997): "Maximum Likelihood for Blind Separation and Deconvolution of Noisy Signals Using Mixture Models," Proc. ICASSP'97 Munich, vol. 5, 3617-20.

[36] PAL, M. (1980): "Consistent Moment Estimators of Regression Coefficients in the Presence of Errors-in-Variables," *Journal of Econometrics*, 14, 349-364.

[37] REIERSOL, O. (1950): "Identifiability of a Linear Relation Between Variables which are Subject to Error," *Econometrica*, 9, 1-24.

[38] ROBIN, J.M., and R.J. SMITH (2000): "Tests of Rank," *Econometric Theory*, 16, 151-175.

[39] ROSS, S.A. (1976): "The Arbitrage Pricing Theory of Capital Asset Pricing," *Journal of Economic Theory*, 16, 341-360.

[40] SPEARMAN, C. (1904): "General intelligence, objectively determined and measured," *American Journal of Psychology*, 15, 201-293.

[41] STEGEMAN, A., and A. MOOIJAART (2007): "Independent Factor Analysis by Least Squares," *mimeo*.

[42] THURSTONE, L.L. (1947): *Multiple-Factor Analysis*, University of Chicago Press, Chicago.